FHS QE Notes

Harry Folkard, Keble College

2022

Abstract

These are my QE notes made for my finals in 2022. They cover all of the topics. In my opinion these notes miss out a fair amount of useful information. This is because I took econometrics and hence focussed most of my revision there, just brushing up on the bits of QE that I needed for the exam. For much more detail please check my micro- and macro- econometrics notes. Nonetheless feel free to use these notes and pass them on to others. Please note, however, that these have just been made by a student and not checked over. They likely contain errors, so it will be worth checking things for yourself. Thanks to Kevin Sheppard, James Duffy and Vanessa Berenguer Rico - these notes are just my interpretation of their lectures and tutorials.

Contents

Nonstationary Times Series **58**

Unit Roots & Stochastic Trends **60**

Orders of Integration **65**

Spurious Regression vs Co-integration **66**

Dynamic Causal Effects **68**

Probability & Statistics

Definitions

Probability Space

Denoted (Ω, A, P) is a set Ω, a (nonempty) collection of subsets A, and a probability function (measure) P defined on A.

Conditional probability

$$P(A \mid B) = \frac{P(A\ B)}{P(B)}$$

Total Probability Theorem

If $\{E(1), \ldots, E(r)\}$ is a partition of Ω such that $P(E(i)) > 0$ for all i, then,

$$P(A) = \sum_{i=1}^{r} P(AnE(i)) = \sum_{i=1}^{r} P(A \mid E(i))P(E(i))$$

- A partition is a collection of disjoint (mutually exclusive) events.
- An Example:
  - There are two Urns (1) and (2)
  - (1) has 3 white and 7 red balls, (2) has 6 white and 4 red balls
  - Heads = take from (1), tails = take from (2)
  - P(ball is white) = P(white n (1)) + P(white n (2))
  - P(ball is white) = P(white | (1))P((1)) + P(white | (2))P((2))
  - P(white) = 3/10 x ½ + 6/10 x ½ = 9/20

Bayes Theorem

This is very easily derivable from conditional probability.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Probability Mass Function

$$f_X(x_i) = P(X = x_i)\ ,\ \sum_i P(X = x_i) = 1$$

Cumulative Distribution Function

$$F_X(x_i) = P(X \leq x_i)$$

Probability Density Function

$$P(a \leq X \leq b) = \int_{-a}^{b} f_X(y)dy\ \ ,\ \ \int_{-\infty}^{\infty} f_X(y)dy = 1$$

Cumulative Distribution Function

$$F_x(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(y)dy\ \ ,\ \ -\infty \leq x \leq \infty$$

Key Functions

Expected Value

$$E[X] = \sum_{i=1}^{...} x_i P(X = x_i) = ...$$

Expected Value is a measure of *centrality*, with the properties,

$$E[a + bX + cY] = a + bE[X] + cE[Y]$$

Variance

$$Var[X] = \sum_{i=1}^{k} \{x_i - E[X]\}^2 P(X = x_i) = ...$$
$$Var[X] = E[\{X - E[X]\}^2] = E[X^2] - E[X]^2$$

The variance is a measure of *dispersion*, with the properties,

$$\begin{aligned} Var[a + bX] &= b^2 Var[X] \\ Var[Y + X] &= Var[Y] + Var[X] + 2Cov[Y, X] \\ Var[Y - X] &= Var[Y] + Var[X] - 2Cov[Y, X] \\ Var[\sum_{i=1}^{n} X_i] &= \sum_{i=1}^{n} Var[X_i] + 2\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Cov[X_i, X_j] \end{aligned}$$

Covariance

$$Cov[X, Y] = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[y]$$

With the properties,

$$\begin{aligned} Cov[aX + b, cW + dV] &= acCov[X, W] + adCov[X, V] + cCov[b, W] + eCov[b, V] \\ &= acCov[X, W] + adCov[X, V] \end{aligned}$$

Correlation Coefficient

$$Corr[X, Y] = \frac{Cov[X, Y]}{\sqrt{Var[Y]Var[X]}}$$

Skewness

$$E[\{\frac{X - E[X]}{\sigma}\}^3]$$

Kurtosis

$$E[\{\frac{X - E[X]}{\sigma}\}^4]$$

Key Distributions

Normal Distribution

$$X \sim N(\mu, \sigma^2)$$

Standardising:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

Calculating Probabilities:

Φ gives the standard normal CDF, while ϕ gives the standard normal PDF.

$$\begin{aligned}
&P(X \leq x) = P(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}) = P(Z \leq \frac{x - \mu}{\sigma}) = \Phi(\frac{x - \mu}{\sigma}) \\
&P(Z \leq c_1) = \Phi(c_1) \\
&P(Z \geq c_2) = 1 - \Phi(c_2) \\
&P(c_1 \leq Z \leq c_2) = \Phi(c_2) - \Phi(c_1)
\end{aligned}$$

If X and Y are bivariate normal then Y and X are uncorrelated iff Y and X are independent.

Uniform Distribution

Density function:

$$f_X(x) = \frac{1}{b - a} \quad \text{, for } a < x < b$$

CDF:

$$F_X(x) = \begin{cases} 0 & , \quad -\infty < x < a \\ \frac{x-a}{b-a} & , \quad a \leq x < b \\ 1 & , \quad b \leq x < \infty \end{cases}$$

Bernoulli

Probability mass function:

$$f_X(x) = P(X = x) = \begin{cases} p & , \text{ if } x = 1 \\ (1 - p) & , \text{ if } x = 0 \\ 0 & , \text{ otherwise} \end{cases}$$

CDF:

$$F_X(x) = P(X \leq x) = \begin{cases} 0 & , \quad -\infty < x < 0 \\ (1 - p) & , \quad 0 \leq x < 1 \\ 1 & , \quad 1 \leq x < \infty \end{cases}$$

Binomial

Probability mass function:

$$f_X(x) = P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \text{ for } x = 0, 1, 2, 3, ..., n$$

Moments:

- Expectation is just given by summing Bernoulli RV's, hence
$$E[X] = np$$
- Variance is again the sum of Bernoulli RV's, hence
$$Var(X) = np(1 - p)$$

Sample & Asymptotic Properties

Sample Properties

$X_i \sim iid(\mu, \sigma^2)$

Sample Mean: $\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- Expectation:
$$E[\bar{X_n}] = E[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu$$
- Variance:
$$Var(\bar{X_n}) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = (\frac{1}{n})^2[\sum_{i=1}^{n} Var(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Cov(X_i, X_j)] = \frac{1}{n^2}n\sigma^2 + \frac{1}{n^2}0 = \frac{\sigma^2}{n}$$
- Standard Error (*Standard Deviation*) and Estimated Standard Error (*Standard Error*):
$$s.d.(\bar{X}_n) = \frac{\sigma}{\sqrt{n}} \quad s.e.(\bar{X}_n) = \frac{\hat{\sigma}_n}{\sqrt{n}}$$

Sample Variance: $V\hat{a}r(X) = \hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

Sample Covariance: $C\hat{o}v(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)$

Law of Large Numbers

Theorem *(Law of Large Numbers by Chebyshev)*

For $i = 1, ..., n$ let x_i be independent and identically distributed with finite mean, μ, and variance σ^2. Then, as $n \to \infty$,
$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow{P} \mu$$

Central Limit Theorems

Theorem *(Central Limit Theorem by Lindeberg-Levy)*

For $i = 1, ..., n$ let x_i be independent and identically distributed with finite mean, μ, and variance σ^2. Then, as $n \to \infty$,
$$\frac{\sqrt{n}(\bar{x}_n - \mu)}{\sigma} \xrightarrow{D} N(0, 1)$$

Point Estimation

An **estimator** is a function of a sample data to be drawn randomly from the population; it is a random variable. For example the sample mean or sample variance.

An **estimate** is the numerical value of the estimator when a specific sample is drawn; it is non-random

Confidence Interval Estimation

We observe the sample mean $\bar{X}_n$, a CI gives all the values of μ that are supported by the data.

Having observed $\bar{X}_n$ a 95% CI is given by all the possible μ's that are supported by the data we have collected,

$$95\%\ CI : \bar{X}_n \pm 1.96 \times s.e.(\bar{X}_n)\ \ ,\ \text{where } X_i \sim iid(\mu, \sigma^2)\ ,\ s.e.(\bar{X}_n) = \frac{\hat{\sigma}_n}{\sqrt{n}}$$

Interpretations

Interpretation (1): Interval Estimation, so it gives a measure of the uncertainty of the point estimate $\bar{X}_n$ (which values of μ are supported by the data).

Re-interpretation (1): A CI is the set of non-rejected null hypothesis.

Interpretation (2): If you were to conduct this experiment 100 times, in 95% of the corresponding samples you will find that $\mu \in CI$

Hypothesis Testing

Null and Alternative Hypotheses

$$H_0 : \mu = \mu_0\ ,\ H_1 : \mu < \mu_0$$

Test-statistic and Distribution

$$t_n = \frac{\bar{X}_n - \mu_0}{s.e.(\bar{X}_n)} \sim N(0,1)\ \ [\text{Under } H_0]$$

$$\text{Where } s.e.(\bar{X}_n) = \frac{\hat{\sigma}_n}{\sqrt{n}}$$

Decision Rule

- Type I Error: reject the null when it is true.
  - We call type I error α, or the Significance level – i.e. 5% significant implies a 5% probability I will reject the null when it is true.
- Type II Error: do not reject the null when the alternative is true
  - We call the type II error β, and 1˘β = power of the test.

P-value

What is the probability under H_0 of finding evidence against the null beyond the observed t-statistic?

$$p = P(t_n < t_n^{obs} \mid H_0) = \Phi(t_n^{obs})$$

- This is only true for the alternative hypothesis at the top, if the alternative hypothesis were that we think the mean is greater than the value we are using, that is $H_1 : \mu > \mu_0$, then

$$p = P(t_n > t_n^{obs} \mid H_0) = 1 - \Phi(t_n^{obs})$$

- Also, this is only true for a one tailed test, if we were doing a two tailed test then

$$p = 2P(t_n > |t_n^{obs}| \mid H_0) = 2(1 - P(t_n < |t_n^{obs}| \mid H_0)) = 2(1 - \Phi(|t_n^{obs}|))$$

One-tailed or Two-tailed?

One-tailed or Two-tailed?

- One sided is more powerful
- But with a one-sided test we miss the other side of the distribution, and therefore opportunity to reject the null.
- Hence we should only use a one-tailed test if we have good reason to do so.

E.g. JTP: Job Training Programme

If wages fall then you don't really care, You'll only implement the scheme if wages rise significantly, hence can ignore the bottom half of the distribution.

(If it is not a good programme, it doesn't matter how not bad it is – if it's bad then we don't implement it, regardless of the level of badness)

Bivariate Statistics

Hypothesis Testing if two means are equal,

E.g. Is mean pay the same for men and women?

$$H_0 : \mu_w = \mu_m \Leftrightarrow \mu_w - \mu_m = 0$$
$$H_1 : \mu_w \neq \mu_m \Leftrightarrow \mu_w - \mu_m \neq 0$$

$$t_n = \frac{\bar{X}_{w,n} - \bar{X}_{m,n} - 0}{s.e.(\bar{X}_{w,n} - \bar{X}_{m,n})} = \frac{\bar{X}_{w,n} - \bar{X}_{m,n}}{\sqrt{\frac{\hat{\sigma}^2_{w,n}}{n} + \frac{\hat{\sigma}^2_{m,n}}{n}}}$$

This assumes that the ***two samples are independent***.

Binomial Testing

Suppose $X \sim B(n, p)$. When n is large then $X \sim N(np, np(1-p))$. This is known as the **binomial approximation to the normal**.

Test statistics:

$$Z = \frac{X - \mu}{\sigma} = \frac{X - np}{\sqrt{np(1-p)}}$$

Or we could consider the mean number of success, rather than the total number, hence,

$$\mu = p\ ,\ \sigma = \sqrt{p(1-p)}$$
$$\mu_{\bar{X}} = p\ ,\ \sigma_{\bar{X}} = \sqrt{\frac{p(1-p)}{n}}$$

Hence,

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Linear Regression

Causal, Population, and Sample Model

Causal Model	Population Model	Sample (OLS) Linear Regression
$Y = \beta_0 + \beta_1 X + u$	$Y = b_0^* + b_1^* X + e$	$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$

Where,
Y = dependent variable
X = independent variable
u collects everything else relevant to the determination of Y other than X.
	With $E[e] = 0$, $E[Xe] = 0$
	The population linear regression asks,
	'what is the best linear predictor of Y using only X?'
	The answer to this question is given by solving,
	Min $E[Y - (b_0 + b_1 X)]^2$
	FOCs,
	$0 = E[Y - b_0^* - b_1^* X]$
	$0 = E[Y - b_0^* - b_1^* X]X$
	Solution,
	$b_1^* = \frac{Cov(X,Y)}{Var(X)}$
	$b_0^* = E[Y] - b_1^* E[X]$
		With $\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i = 0$, $\frac{1}{n}\sum_{i=1}^{n} X_i \hat{u}_i = 0$
		$(\hat{\beta}_0, \hat{\beta}_1)$ are found by solving,
		$(\hat{\beta}_0, \hat{\beta}_1) = argmin\ \frac{1}{n}\sum_{i=1}^{n}[Y_i - \beta_0 - \beta_1 X]^2$
		Solution,
		$\hat{\beta}_1 = \frac{C\hat{o}v(Y,X)}{V\hat{a}r} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$
		$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

u is orthogonal to X iff $E[u] = 0$, $E[Xu] = 0$, since this implies X and u are uncorrelated.	Define the regression error e as, $e := Y - b_0^* - b_1^* X$ By the FOCs the error is orthogonal to X. $E[e] = E[Y - b_0^* - b_1^* X] = 0$ $E[Xe] = E[Y - b_0^* - b_1^* X]X = 0$ We say in this model e and X are orthogonal by construction.	The linear regression function provides a linear approximation of the conditional mean $E[Y \mid X = x_1] \approx \hat{\beta}_0 + \hat{\beta}_1 x_1$ When the conditional expectation function (CEF) is linear the linear regression function is the best approximation of it simplicter
If the causal model u satisfies OR then the population model and causal model coincide, hence OLS estimates can be causally interpreted.	Under OR the causal and population models coincide.	The sample regression always consistently estimates the population parameters. Under OR the population and causal models coincide and hence the sample regression also consistently estimates the causal model

So we have the population model, which tells us what the best linear predictor of Y using X is, and for which OR holds by construction. The sample linear regression always consistently estimates the population regression. When OR holds in the causal model, then the causal model coincides with the population model and hence the sample regression consistently estimates the causal model.

OR, MI, and IN

OR Orthogonal (and mean zero):

$$\begin{aligned}&E[u] = 0 \text{ and } E[Xu] = 0\ ,\ \text{or equivalently,}\\&E[u] = 0 \text{ and } Cov(X, u) = 0\end{aligned}$$

MI Mean independent (and mean zero):

$$E[u] = 0 \text{ and } E[u \mid X] = E[u]$$

IN Independent (and mean zero):

$$E[u] = 0 \text{ and } u \perp\!\!\!\perp X$$

IN ⇒ MI ⇒ OR

If the causal model error satisfies OR then the causal model coincides with the linear regression model.

Mean Zero Error

Interestingly, our error mean zero assumption is actually unnecessarily strong. We are able to recover $(\hat{\beta}_0, \hat{\beta}_1)$ without it.

Suppose OR such that that $Cov(X, u) = 0$ but that $E[u] \neq 0$.

From our FOCs we know that,

$$\begin{aligned}&(1)\ E[Y - \beta_0 - \beta_1 X] = E[u]\\&(2)\ E[(Y - \beta_0 - \beta_1 X)X] = E[uX]\end{aligned}$$

To recover,

$$\begin{aligned}(2)\ &E[(Y - \beta_0 - \beta_1 X)X] = E[uX]\\&E[YX] - \beta_0 E[X] - \beta_1 E[X^2] = E[uX]\\(1)\ &E[Y - \beta_0 - \beta_1 X] = E[u]\\&E[Y] - \beta_0 - \beta_1 E[X] = E[u]\end{aligned}$$

Immediately we can see that

$$\beta_0 = E[Y] - \beta_1 E[X] - E[u]$$

Using this further,

$$\begin{aligned}&E[YX] - \beta_0 E[X] - \beta_1 E[X^2] = E[uX]\\&E[YX] - (E[Y] - \beta_1 E[X] - E[u])E[X] - \beta_1 E[X^2] = E[uX]\\&E[YX] + \beta_1 (E[X])^2 - \beta_1 E[X^2] - E[Y]E[X] - E[u]E[X] = E[uX]\\&E[YX] - E[Y]E[X] - E[u]E[X] - E[uX] = \beta_1\{E[X^2] - E[X]^2\}\\&Cov(Y, X) - Cov(u, X) = \beta_1 Var(X)\\&Cov(Y, X) - 0 = \beta_1 Var(X)\end{aligned}$$

And so finally,

$$\begin{aligned}\beta_1 &= \frac{Cov(Y, X)}{Var(X)}\\\beta_0 &= E[Y] - \beta_1 E[X] - E[u]\end{aligned}$$

The point being that we can recover β_1 without it being true that $E[u] = 0$, and infact providing we can estimate $E[u]$ we can recover β_0 as well.

Causal/Descriptive Distinction

Descriptive Interpretation	Causal Interpretation
How is X correlated with Y. Think ‘Are’. Example: Are test scores higher in schools with smaller classes? If OR fails, then we have a descriptive interpretation. That is ’$\hat{\beta}_1$ gives, all else equal, how, on average, Y changes with X’. Example: ’Tests scores are, on average, 2.28 higher in districts with 1 less student per teacher.’	What would happen to Y if we were to change X. Think ‘Would’. *Example:* Would a reduction in class sizes improve test scores? If OR holds, then we have a causal interpretation. That is ’$\hat{\beta}_1$ gives, all else equal, the causal effect of X on Y’. Example: ’If we were to reduce class sizes by 1, test scores would, all else equal, increase on average by 2.28.’

Regression and Conditional Expectation

Population regression provides the best linear predication of Y given that we observe $X = x$

$$min_{b_0,b_1} \ E[Y - (b_0 + b_1 X)^2]$$

The conditional expectation of Y given X gives the best predication of Y among all possible functions of X, including non-linear ones,

$$\begin{aligned} min_m \ & E[Y - m(X)^2] \\ & E[Y - m(X)^2] = E[(Y - E[Y \mid X]) - (m(X) - E[Y \mid X])^2] \\ & \qquad = E[\epsilon + g(X)^2] \text{ where } \epsilon := Y - E[Y \mid X] \text{ and } g(X) := -(m(X) - E[Y \mid X]) \\ & \qquad = E[\epsilon + g(X)^2] = E[\epsilon^2] + 2E[g(X)\epsilon] + E[g(X)^2] \\ & \qquad \text{notice that } E[g(X)\epsilon] = E[g(X)E[\epsilon \mid X]] = E[Y \mid X] - E[Y \mid X] = 0 \\ & \qquad = E[\epsilon + g(X)^2] = E[\epsilon^2] + E[g(X)^2] \end{aligned}$$

Which is minimised when $g(X) = 0,$ hence $m(X) = E[Y \mid X]$

When the CEF (conditional expectation function) is linear then regression provides the best approximation of the CEF and hence both provide the same solution to the minimisation problem.

Otherwise regression provides a linear approximation of the CEF.

Multivariate Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

Where now OR is given by,

OR Orthogonal to X_1 and X_2 (and mean zero):

$$\begin{aligned} &E[u] = 0\ ,\ E[X_1 u] = 0,\ \text{and}\ E[X_2 u] = 0 \text{ or equivalently,} \\ &E[u] = 0\ ,\ Cov(X_1, u) = 0,\ \text{and}\ Cov(X_2, u) = 0 \end{aligned}$$

MI Mean independent of X_1 and X_2 (and mean zero):

$$E[u \mid X_1, X_2] = 0$$

We might have to/want to use a proxy variable 'Z' in a multivariate regression in order to predict X when we are unable to measure X. Estimator on a Proxy is always considered descriptively.

Conditional mean is as it was the case with the simple regression model, the multivariate linear regression function provides a linear approximation to the conditional (or an exact approximation if the conditional mean is linear).

$$\begin{aligned} &E[Y \mid X_1 = x_1, ..., X_k = x_k] \approx \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k \\ &\beta_1 = \frac{\partial}{\partial x_1} E[Y \mid X_1 = x_1, ..., X_k = x_k] \end{aligned}$$

Frisch-Waugh-Lovell Theorem

FWL theorem explains the mechanics of multivariate regression.

Our problem is, as ever,

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \\ (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) &= argmin\ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i})^2 \end{aligned}$$

We can solve this by setting up the regression,

$$\begin{aligned} X_1 &= \pi_0 + \pi_2 X_2 + \tilde{X}_1 \\ &\text{with } E[\tilde{X}_1] = 0\ ,\ E[X_2 \tilde{X}_1] = 0 \end{aligned}$$

Here $\tilde{X}_1$ collects the part of X_1 that is uncorrelated with X_2. (In the proof this is a population linear regression hence OR holds by construction)

We then set up the regression,

$$\begin{aligned} Y &= \beta_0 + \beta_1(\pi_0 + \pi_2 X_2 + \tilde{X}_1) + \beta_2 X_2 + u \\ &= \beta_0 + \beta_1 \pi_0 + \beta_1 \tilde{X}_1 + (\beta_1 \pi_2 + \beta_2) X_2 + u \\ &= \gamma_0 + \beta_1 \tilde{X}_1 + \gamma_2 X_2 + u \\ &\text{let } \gamma_2 X_2 + u = \epsilon, \text{ by the definition of } \tilde{X}_1 \text{ we then know that } Cov(\epsilon, \tilde{X}_1) = 0 \end{aligned}$$

Hence we can regress Y on $\tilde{X}_1$ alone, giving the solution,

$$\beta_1 = \frac{Cov(Y, \tilde{X}_1)}{Var(\tilde{X}_1)}$$

In the multivariate case,

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + u$$

$$\beta_1 = \frac{Cov(Y, \tilde{X}_1)}{Var(\tilde{X}_1)} \text{ where } X_1 = \pi_0 + \pi_2 X_2 + ... + \pi_k X_k + \tilde{X}_1$$

Providing OR holds, the OLS estimator satisfies the sample analogue,

$$\hat{\beta}_1 = \frac{C\hat{o}v(Y, \tilde{X}_1^{"})}{V\hat{a}r(\tilde{X}_1^{"})} \text{ where } X_1 = \hat{\pi}_0 + \hat{\pi}_2 X_2 + ... + \hat{\pi}_k X_k + \tilde{X}_1^{"}$$

Perfect Multicollinearity

Perfect Multicollinearity describes a situation in which multivariate linear regression and FWL theorem fail.

The more X_1 is explained by X_2,. . . ,X_k, the smaller is, since is the part of X_1 uncorrelated with X_2,. . . ,X_k.

If X_2,. . . ,X_k perfectly explain X_1 then $\tilde{X}_1 = 0$ hence the regression fails.

Example: Dummy variables

There are 3 types of school district: city, town, and rural.

If $X_1 =$ city, $X_2 =$ town, $X_3 =$ rural, then X_2 and X_3 perfectly explain X_1, hence $\tilde{X}_1 = 0$ and the regression fails.

Inference

Measures of fit

These tell us how well Y is explained by the model. That is how well Y is explained by $\beta_0 + \sum_{i=1}^{k} \beta_i X_i$.

The better explained Y is by the model, the better the fit of the regression

Standard error of regression (SER)

The variance of error is given by,

$$\sigma_u^2 = Var(u_i) = E[u_i^2] - E[u_i]^2 = E[u_i^2]$$

And measures the extent to which Y_i departs from the regression line.

$\sigma_u^2 = 0$ iff Y_i always lies exactly on the regression line.

This is estimated by,

$$s_{\hat{u}}^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} \hat{u}_i^2$$

Note that $\sum_{i=1}^{n} (\hat{u}_i - \bar{\hat{u}}_i)^2 = \sum_{i=1}^{n} \hat{u}_i^2$ since $\sum_{i=1}^{n} \hat{u}_i = 0$ from the FOCs.

Hence the SER is given by,

$$\text{SER} : s_{\hat{u}} = \sqrt{\frac{1}{n-k-1} \sum_{i=1}^{n} \hat{u}_i^2}$$

Regression R-squared

R-squared is defined as the fraction of the variability of Y_i that is explained by the model [explained error / total error]

More regressors never decreases R-squared, and almost always increases it.

$$\text{TSS} := \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$
$$\text{ESS} := \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$
$$\text{SSR} := \sum_{i=1}^{n} \hat{u}_i^2$$

$$R^2 := \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{SSR}}{\text{TSS}}$$

Where,

- TSS : Total Sum of Squares,
- ESS : Explained Sum of Squares, and
- SSR : Sum of Square Residuals.

It is also useful to realise that,

$$\begin{aligned}\text{TSS} &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}((\hat{Y}_i + \hat{u}_i) - \bar{Y})^2 \\ &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}\hat{u}_i^2 - 2\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})\hat{u}_i \\ &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}\hat{u}_i^2 \\ &= \text{ESS} + \text{SSR}\end{aligned}$$

Adjusted R-squared

$$\bar{R}^2 : 1 - \frac{n-1}{n-k-1}\frac{\text{SSR}}{\text{TSS}} = 1 - \frac{\text{SSR}/(n-k-1)}{\text{TSS}/(n-1)} = 1 - \frac{s_{\hat{u}}^2}{s_y^2}$$

Inference on Regression Parameters

We want to know the answer to these four questions,

(1) Is $\hat{\beta}_1$ a good estimator?
(2) What is the distribution of $\hat{\beta}_1$ in large samples?
(3) How can we quantify the uncertainty associated with $\hat{\beta}_1$?
(4) How can we test hypotheses about $\hat{\beta}_1$?

Assumptions we will use,

(i) u_i satisfies OR.
(ii) Y_i, X_{1i},. . . , X_{ki} are iid.
(iii) Large outliers are unlikely.
(iv) No perfect multicollinearity.

(1) Is $\hat{\beta}_1$ a good estimator? And (2) What is the distribution of $\hat{\beta}_1$ in large samples?

In short yes $\hat{\beta}_1$ is a good estimator. This is the case because $\hat{\beta}_1$ is,

1. Unbiased
$$E[\hat{\beta}_1] = \beta_1$$
2. Consistent
$$\hat{\beta}_1 \xrightarrow{P} \beta_1$$
3. Asymptotically normal
$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{D} N(0, \omega^2_{\beta_1})$$
4. Efficient (smallest variance amongst linear unbiased estimators) - BLUE : Best Linear Unbiased Estimator

(3) How can we quantify the uncertainty associated with $\hat{\beta}_1$?

With the Standard Error! Consider first the asymptotic distribution and hence the asymptotic variance,

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{D} N(0, \omega^2_{\beta_1})$$

Which implies that,

$$\hat{\beta}_1 \xrightarrow{D} N(\beta_1, \frac{1}{n}\omega^2_{\beta_1})$$

Hence the standard error of $\hat{\beta}_1$,

$$s.e.(\hat{\beta}_1) = \frac{\hat{\omega}_{\beta_1}}{\sqrt{n}}$$

(A) Under **Heteroskedasticity**

The conditional variance of u does depend on the regressors – variance changes with the regressors.

We know that in this case the asymptotic variance of $\hat{\beta}_1$ is given by,

$$\omega^2_{\beta_1,hetero} = \frac{Var(\tilde{X}_1 u)}{[Var(\tilde{X}_1)]^2}$$

(B) Under **Homoskedasticity**

Where the conditional variance of u does not depend on the regressors.

The asymptotic variance of $\hat{\beta}_1$ in this case is given by,

$$\omega^2_{\beta_1,homo} = \frac{\sigma_u^2}{Var(\tilde{X}_1)}$$

Why?

Because the heteroskedastic asymptotic variance is,

$$\omega^2_{\beta_1,hetero} = \frac{Var(\tilde{X}_1 u)}{[Var(\tilde{X}_1)]^2} = \frac{E[\tilde{X}_1^2 u^2]}{(E[\tilde{X}_1^2])^2}$$

But we know that under homoskedasticity,

$$E[\tilde{X}_1^2 u^2] = E[E[\tilde{X}_1^2 u^2 \mid X]] = E[\tilde{X}_1^2 E[u^2 \mid X]] = \sigma_u^2 E[\tilde{X}_1^2]$$

Hence,

$$\begin{aligned}\omega^2_{\beta_1,homo} &= \frac{Var(\tilde{X}_1 u)}{[Var(\tilde{X}_1)]^2} = \frac{E[\tilde{X}_1^2 u^2]}{(E[\tilde{X}_1^2])^2} = \frac{\sigma_u^2 E[\tilde{X}_1^2]}{(E[\tilde{X}_1^2])^2} \\ &= \frac{\sigma_u^2}{E[\tilde{X}_1^2]} \\ &= \frac{\sigma_u^2}{Var(\tilde{X}_1)}\end{aligned}$$

IMPORTANT TAKEAWAY – Standard errors are different for hetero/homo-skedasticity.

$$(homo)\ s.e.(\hat{\beta}_1) = \frac{1}{\sqrt{n}} \frac{s_{\hat{u}}}{s.\hat{d}.(\tilde{X}_1^{"})}$$

$$(hetero)\ s.e.(\hat{\beta}_1) = \frac{1}{\sqrt{n}} \frac{s.\hat{d}.(\tilde{X}_1^{"}\hat{u})}{V\hat{a}r(\tilde{X}_1^{"})}$$

(4) How can we test hypotheses about $\hat{\beta}_1$?

1. t-tests

   $$H_0 : \beta_1 = b\ ,\ H_1 : \beta_1 \neq b$$

   $$t(b) = \frac{\hat{\beta}_1 - b}{s.e.(\hat{\beta}_1)} \sim N(0,1) \text{ (Under } H_0)$$

   Then,

   $$\text{Reject } H_0 \text{ if } |t(b)| > c$$

2. p-values

   $$H_0 : \beta_1 = b\ ,\ H_1 : \beta_1 \neq b$$

   $$t(b) = \frac{\hat{\beta}_1 - b}{s.e.(\hat{\beta}_1)} \sim N(0,1) \text{ (Under } H_0)$$

   Our sample delivers the t-value t^{act}

   $$p = P(|t(b)| > |t^{act} = 2\Phi(-|t^{act}|)$$

   Then,

   $$\text{Reject } H_0 \text{ if } p > c$$

3. Confidence Intervals

   $$C = \{\hat{\beta}_1 \pm c_\alpha \times s.e.(\hat{\beta}_1)\}$$

Testing Multiple Hypotheses: The F-test

Suppose the model,

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_q X_{qi} + ... + \beta_k X_{ki} + u_i$$

And that we want to test q restrictions,

$$\begin{aligned} &H_0 : \beta_1 = \beta_2 = ... = \beta_q = 0 \\ &H_1 : \beta_I \neq 0 \;\; \exists I \in \{1, ...q\} \end{aligned}$$

We start by developing two models:

(1) Unrestricted Model:

$$\begin{aligned} &Y_i = \hat{\beta}_{0,un} + \hat{\beta}_{1,un} X_{1i} + ... + \hat{\beta}_{q,un} X_{qi} + ... + \hat{\beta}_{k,un} X_{ki} + \hat{u}_{i,un} \\ &SSR_{un} = \sum_{i=1}^{n} \hat{u}_{i,un} \end{aligned}$$

(2) Restricted Model:

$$\begin{aligned} &Y_i = \hat{\beta}_{0,rs} + \hat{\beta}_{q+1,rs} X_{qi} + ... + \hat{\beta}_{k,rs} X_{ki} + \hat{u}_{i,rs} \\ &SSR_{rs} = \sum_{i=1}^{n} \hat{u}_{i,rs} \end{aligned}$$

The test,

$$F = \frac{SSR_{rs} - SSR_{un}}{SSR_{un}} \frac{n-k-1}{q} = \frac{(SSR_{rs} - SSR_{un})/q}{SSR_{un}/(n-k-1)} \xrightarrow{D} F_{q,\infty}$$

Where, to be clear, q - number of restrictions; n - sample size; k - number of regressors.

Decision rule,

$$\text{Reject } H_0 \text{ if } F > c_\alpha$$

We can use the F-test to see if the model is at all relevant. Simply test the null hypothesis that none of the variables are relevant ($\forall k, \;\beta_k = 0$). If we reject this hypothesis then the regression is not useless.

General F-test

Suppose the model,

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_q X_{qi} + ... + \beta_k X_{ki} + u_i$$

And that we want to test $q = 2$ restrictions,

$$\begin{aligned} &H_0 : \beta_1 = b \text{ and } \beta_2 + \beta_3 = c \\ &H_1 : \beta_1 \neq 0 \text{ or } \beta_2 + \beta_3 \neq c \end{aligned}$$

(1) Unrestricted Model:

$$\begin{aligned} &Y_i = \hat{\beta}_{0,un} + \hat{\beta}_{1,un} X_{1i} + ... + \hat{\beta}_{q,un} X_{qi} + ... + \hat{\beta}_{k,un} X_{ki} + \hat{u}_{i,un} \\ &SSR_{un} = \sum_{i=1}^{n} \hat{u}_{i,un} \end{aligned}$$

(2) Restricted Model:

$$Y_i = \hat{\beta}_{0,rs} + bX_{1i} + (c - \hat{\beta}_{3,rs})X_{2i} + \hat{\beta}_{3,rs}X_{3i} + ... + \hat{\beta}_{k,rs}X_{ki} + \hat{u}_{i,rs}$$

Notice he model is restricted by making a substitution for $\hat{\beta}_1 = b$ and $\hat{\beta}_2 + \hat{\beta}_3 = c$. Then take any variables with known coefficients to the LHS,

$$\begin{aligned} Y_i - bX_{1i} - cX_{2i} &= \hat{\beta}_{0,rs} + \hat{\beta}_{3,rs}(X_{3i} - X_{2i}) + ... + \hat{\beta}_{k,rs}X_{ki} + \hat{u}_{i,rs} \\ Y_i^* &= \hat{\beta}_{0,rs} + \hat{\beta}_{3,rs}(X_{3i} - X_{2i}) + ... + \hat{\beta}_{k,rs}X_{ki} + \hat{u}_{i,rs} \\ SSR_{rs} &= \sum_{i=1}^{n} \hat{u}_{i,rs}^2 \end{aligned}$$

The test,

$$F = \frac{SSR_{rs} - SSR_{un}}{SSR_{un}} \frac{n-k-1}{q} = \frac{(SSR_{rs} - SSR_{un})/q}{SSR_{un}/(n-k-1)} \xrightarrow{D} F_{q,\infty}$$

Where, to be clear, q - number of restrictions; n - sample size; k - number of regressors.

Decision rule,

$$\text{Reject } H_0 \text{ if } F > c_\alpha$$

Non-Linearities

- Models will **always** be linear in parameters for QE
- Perfect multicollinearity doesn't matter here because it only matters if the variables **linearly** explain one another.

Polynomials

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_r X^r + u$$

We can test how many polynomials we need by using the F-test to test the null of linearity against the r^{th} degree polynomial. The r at which we reject the null is how many we need,

$$H_0 : \beta_2 = ... = \beta_r = 0$$

Suppose $r = 2$ For small changes the causal effect of X on Y is given by,

$$\frac{\partial Y}{\partial X} = \beta_1 + 2\beta_2 X$$

For larger changes the causal effect is calculated by the difference between the regression functions, i.e.

$$\Delta Y = Y(X + \delta X) - Y(X)$$

Statistical Inference on Marginal Effect

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + u$$

Marginal effect:

$$\frac{\partial Y}{\partial X}|_{X=x} = Y'(x) = \beta_1 + 2\beta_2 x$$

Test:

$$H_0 : Y'(x) = b$$
$$H_1 : Y'(x) \neq b$$

Notice that $H_0 : \beta_1 + 2\beta^2 x - b = 0$

Then use the general F-test described above, or since here $q = 1$ (one restriction), a t-test,

$$t(b) = \frac{\hat{Y}'(x) - b}{s.e.(\hat{Y}'(x))} = \frac{\hat{\beta}_1 + 2\hat{\beta}^2 x - b}{s.e.(\hat{\beta}_1 + 2\hat{\beta}^2 x)}$$

Logarithms

(1) Linear-Log:

$$Y = \beta_0 + \beta_1 log X + u$$
$$\Delta Y = \beta_1 \Delta log X \approx \beta_1 \frac{\Delta X}{X}$$

"A 1% increase in X has a $0.01 \times \beta_1$ effect on Y"

(2) Log-Linear

$$logY = \beta_0 + \beta_1 X + u$$

$$\Delta logY = \beta_1 \Delta X \Rightarrow \frac{\Delta Y}{Y} = \beta_1 \Delta X$$

"A unit increase of X increases Y by $\beta_1 \times 100\%$". Neat rule: $(e^{\hat{\beta}_1} - 1) \times 100$ gives the percentage change in Y given a change in X.

(3) Log-Log

$$logY = \beta_0 + \beta_1 logX + u$$

$$\Delta logY = \beta_1 \Delta logX \Rightarrow \frac{\Delta Y}{Y} = \beta_1 \frac{\Delta X}{X} \Rightarrow \beta_1 = \frac{\Delta Y/Y}{\Delta X/X}$$

"β_1 is the elasticity of Y wrt to X"

Interaction terms

(1) Constant

$$Y = \beta_0 + \beta_1 X + \beta_2 D + u$$

If $D = 1$ then the constant is $\beta_0 + \beta_2$, or β_0 when $D = 0$.

(2) Slope

$$Y = \beta_0 + \beta_1 X + \beta_3 X \cdot D + u$$

Where $\frac{\partial Y}{\partial X} = \beta_1 + \beta_3 D = \begin{cases} \beta_1 \ , \ D = 0 \\ \beta_1 + \beta_3 \ , \ D = 1 \end{cases}$.

(3) Constant & Slope

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 X \cdot D + u$$

In this case the interaction term affects both the constant and the slope.

Problems with Regression Analysis

This section is titled *“(quasi)-experiments and causal effects”* in the notes and lectures, but I didn’t really understand what that title meant.

Exogeneity & Endogeneity

Starting point is again the causal model,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

Will OLS consistently estimate β_1?

- Requires $Cov(X_1, u) = Cov(X_2, u) = 0$
- If this is the case X_1, X_2 are *exogenous.*
- If $Cov(X_1, u) \neq 0$ then X_1 is *endogenous* and OR fails, hence *neither* β_1 nor β_2 is consistently estimated by OLS.

Causes of Endogeneity

(1) Omitted Variables

- Some determinants of Y are ‘buried’ in u and correlated with the X’s - the regressors.
- In this case then OR doesn’t hold. Recall that we need $\forall I \; Cov(X_I, u) = 0$, but if u contains X_J and $\exists I \; Cov(X_I, X_J) \neq 0$ then we have endogeneity.

(2) Measurement Error (in X_1)

- Causes ‘attenuation bias’: $\hat{\beta}_1$ is shrunk towards zero.
- Measurement error in the dependent variable, Y, can be dealt with, though the error increases.

(3) Simultaneity/Reverse Causality

- Y also ‘causes’ X_1 or X_2.
- This topic will be discussed further in the instrumental variables section.

(1) Omitted Variables:

Recall that for our coefficients to have a causal interpretation OR must hold, that is $E[u] = E[X_1 u] = ... = E[X_k u] = 0$ or $E[u] = Cov(X_1, u) = ... = Cov(X_k, u) = 0$. If we omit variables in our regression, then they get caught up in u - they are part of 'everything else' that explains Y. The problem, however, is that now OR might not hold, since the omitted variable(s) could be correlated with one of our regressors.

Omitted Variable Bias (OVB) Formula

Let's define the 'long regression',

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

and assume OR holds, $E[u] = E[uX_1] = E[uX_2] = 0$.

Let's also define the 'short regression',

$$Y = \gamma_0 + \gamma_1 X_1 + \epsilon$$

We know that X_1 in the short regression will be endogenous since $\epsilon = \beta_2 X_2 + u$ and we know that $Cov(X_1, \epsilon) = Cov(X_1, \beta_2 + u) = \beta_2 Cov(X_1, X_2) + Cov(X_1, u)$

In order to be exogenous we would need either:

(1) $\beta_2 = 0 :$ X_2 is irrelevant to Y, or

(2) $Cov(X_1, X_2) = 0 :$ X_2 is uncorrelated with X_1.

Despite these concerns let's do the *population* short regression anyway,

$$Y = \gamma_0 + \gamma_1 X_1 + e$$

where, *by construction*, $E[e] = E[X_1 e] = 0$

We know it will be the case that

$$\gamma_1 = \frac{Cov(Y, X_1)}{Var(X_1)}$$

Using the long regression,

$$\begin{aligned} Cov(Y, X_1) &= Cov(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + u, X_1) \\ &= Cov(\beta_1, X_1) + \beta_1 Cov(X_1, X_1) + \beta_2 Cov(X_2, X_1) + Cov(u, X_1) \\ &= 0 + \beta_1 Var(X_1) + \beta_2 Cov(X_2, X_1) + 0 \\ &= \beta_1 Var(X_1) + \beta_2 Cov(X_2, X_1) \end{aligned}$$

And so,

$$\gamma_1 = \frac{\beta_1 Var(X_1) + \beta_2 Cov(X_2, X_1)}{Var(X_1)} = \beta_1 + \beta_2 \frac{Cov(X_1, X_2)}{Var(X_1)}$$

Notice that $\frac{Cov(X_1,X_2)}{Var(X_1)}$ is the formula for the population of X_2 on X_1. We can define this as π_1.

From this we get the *OVB formula*,

$$\gamma = \beta_1 + \beta_2 \pi_1 \text{ where } X_2 = \pi_0 + \pi_1 X_1 + \tilde{X}_2 \ , \ E[\tilde{X}_2] = E[X_1 \tilde{X}_2] = 0$$

Hence there is no bias ($\gamma = \beta_1$) if:

(A) $\beta_2 = 0 :$ X_2 is irrelevant to the determination of Y.

(B) $\pi_1 = 0 :$ X_1 and X_2 are uncorrelated.

General Model OVB

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + u$$

Where OR is assumed to hold.

If we omitted X_k,

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{k-1} X_{k-1} + e$$

where OR holds by construction.

Hence the *OVB formula*,

$$\gamma_1 = \beta_1 + \beta_k \pi_1$$

Where π_1 is the coefficient on X_1 of the population linear regression of X_k on the other X's $\{1, \dots, (k-1)\}$.

Proxying for Omitted Variables

(A) Single Proxy

Start with $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$ where $E[u] = E[X_1 u] = E[X_2 u] = 0$

Use some proxy for X_2 called W, where W is a valid proxy for X_2 if,

(1) $E[Wu] = 0$

(2) The error e in the (hypothetical) population linear regression $X_2 = \delta_0 + \delta_2 W + e$, $E[u] = E[We] = 0$ satisfies $E[X_1 e] = 0$.

- Error e in the 'best (linear) prediction' made about X_2 on the basis of W alone, isn't correlated with X_1.
- This implies X_1 couldn't help us improve upon this prediction.
- In other words, W must be a sufficiently 'good' predictor of X_2, that what is left over isn't correlated with X_1.
- Doesn't preclude $Cov(X_1, X_2) = 0$, because X_1 may itself be correlated with W.

Where does condition (2) come from?

$$X_2 = \delta_0 + \delta_1 W + e \ , \ E[e] = E[We] = 0$$

Subbing into the model for Y,

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2(\delta_0 + \delta_1 W + e) + u \\ &= \beta_0 + \beta_2\delta_0 + \beta_1 X_1 + \beta_2\delta_1 W + \beta_2 e + u \\ &= (\beta_0 + \beta_2\delta_0) + \beta_1 X_1 + \beta_2\delta_1 W + (\beta_2 e + u) \end{aligned}$$

For OR to hold and us to be able to recover β_1, the condition $Cov(X_1, \beta_2 e + u) = 0$ hence requires that X_1 and e are uncorrelated.

(B) Multiple Proxies

We can proxy for X_2 with more than one proxy, say,

$$X_2 = \delta_0 + \delta_1 W_1 + \delta_2 W_2 + e \ , \ E[e] = E[W_1 e] = E[W_2 e] = 0$$

This really doesn't change much of what we do, other than needing condition (1) to hold for all proxies now, rather than just the one.

(2) Measurement Error:

LHS: Measurement Error in Dependent Variable

Suppose we don't observe Y, but instead we observe Y^*, which is Y with some measurement error $Y^* = Y + e_y$.

If we still want to regress Y on X we can with,

$$\begin{aligned} Y^* - e_y &= \beta_0 + \beta_1 X + u \\ Y^* &= \beta_0 + \beta_1 X + (u + e_y) \end{aligned}$$

The conclusion from this is that, as long as $Cov(X, e_y) = 0$ (the measurement error in Y is not correlated with X) the OLS regression of Y^* on X would consistently estimate β_1.

- Of course we also need OR to hold normally as well so that $Cov(X, u) = 0$, but we are talking about a population regression here not sample, hence that holds by construction.

$Var(\beta_1)$ is now larger due to additional error though.

RHS: Measurement Error in Independent Variable

Suppose now we don't observe X but instead we observe X^*, which is X with some measurement error such that $X^* = X + e_x$

Suppose $Cov(X, e_x) = Cov(Y, e_x) = 0$ - this is a very generous assumption to make!

$$\begin{aligned} Y &= \beta_0 + \beta_1 (X^* + e_x) + u \\ &= \beta_0 + \beta_1 X^* + (\beta_1 e_x + u) \end{aligned}$$

Since $Cov(X^*, e_x) = Cov(X + e_x, e_x) = Var(e_x)$ then, $Cov(X^*, u - \beta_1 e_x) \neq 0$

Hence we cannot consistently estimate β_1.

Could we make it consistent?

$$\begin{aligned} \beta_1^* &= \frac{Cov(Y, X^*)}{Var(X^*)} = \frac{Cov(\beta_0 + \beta_1 X^* + \epsilon, X^*)}{Var(X^*)} = \frac{\beta_1 Var(X^*) + Cov(\epsilon, X^*)}{Var(X^*)} \\ &= \frac{\beta_1 Var(X^*) + Cov(u - \beta_1 e_x, X + e_x)}{Var(X^*)} = \frac{\beta_1 [Var(X^*) + Var(e_x)]}{Var(X^*)} \\ &= \beta_1 [1 - \frac{Var(e_x)}{Var(X) + Var(e_x)}] = \beta_1 [\frac{Var(X)}{Var(X) + Var(e_x)}] \end{aligned}$$

All this maths basically shows that β_1 gets multiplied by something between 0 and 1 and hence β_1^* will shrink towards zero.

So we can't really make it consistent no.

(3) Simultaneity:

Two variables are jointly determined, rather than one being a function of the other.

Easiest place to see this problem is with supply and demand, where,

$$Q = \beta_0 + \beta_1 P + \beta_2 X + u$$

Often X and u are uncorrelated, but the problem is that Q and u are obviously correlated (u gathers things that determine Q).

We also know that Q determines P, when quantity changes price changes, hence P and u are correlated, hence OR doesn't hold.

Ergo, endogeneity.

A Solution: Randomised Control Trials, Natural Experiments, and Heterogeneous Causal Effects

The problem we need to solve is that of $Cov(X,u) \neq 0$, that is the problem of endogeneity - of OR not holding.

Random Control Trials

Randomly assign X to the study participants such that it is independent of their other characteristics.

Hence renders X independent of u by design – which implies $Cov(X,u) = 0$.

Example: OLS Binary Regressor

- $D = 1$: treatment,
- $D = 0$: control
- Difference in means gives us exactly the OLS estimator
$$\frac{Cov(Y,D)}{Var(D)} = E[Y \mid D=1] - E[Y \mid D=0]$$
  - OLS estimator satisfies analogous decomposition,
$$\frac{C\hat{o}v(Y,D)}{V\hat{a}r(D)} = \hat{E}[Y \mid D=1] - \hat{E}[Y \mid D=0]$$
  - Or put more simply,
$$\frac{1}{n_1}\sum_{i \mid D_i=1} Y_i - \frac{1}{n_0}\sum_{i \mid D_i=0} Y_i$$
  - Here the conditional mean is linear, hence OLS exactly coincides with it.

Example: Conditional Random Assignment

- Perfect random assignment is not always possible, since you can't up and move people just for your RCT.
- For example if you are testing the effect of class sizes on test scores, you can't move teachers around the country, but you can randomly assign teachers to classes within a school.
- In this situation you still need to account for school district differences etc, but now standard of teaching is random *within* a school.
- Then just uses the FWL theorem,
$$X = \gamma_0 + \gamma_1 W_1 + ... + \gamma_k W_k + \tilde{X}$$
  - Hence,
$$\beta_1 = \frac{Cov(Y,\tilde{X})}{Var(\tilde{X})}$$

WARNING: Endogenous or Bad Controls

- Only add pre-treatment characteristics to regressions in conditional assignment.
- If you add something that is post treatment then it will be endogenous – a post-treatment characteristic has already been affected by the treatment!!

Heterogeneous Causal Effects

What about if different people have different causal effects?

$$Y_i = \beta_0 + \beta_{1i} X_i + u_i$$

Interestingly nonlinear models like quadratic regression and interaction terms are a case of heterogeneous causal effect since different people do have different causal effects.

$$\beta_1 = \frac{Cov(Y, X)}{Var(X)} = ATE = ACE$$

The "average treatment/causal effect".

Natural Experiments

This is just RCTs but you try and find them naturally.

Example: In Utero Nutrition & Birthweight

- Look at how maternal fasting during Ramadan affects birth weight.

Instrumental Variables

Strategies for Regression so far

$$Y = \beta_0 + \beta_1 X + u$$

Three strategies to estimate the causal effect of X on Y.

(1) Observational data (include proxies, other determinants, etc).

(2) RCTs: Randomly assign X such that $Cov(X, u) = 0$.

(3) Natural experiments: Find settings in which $Cov(X, u) = 0$

If all of these strategies fail then we may need to go beyond regression with instrumental variables.

One Instrument Case

Assumptions

Suppose a variable Z that satisfies the following conditions:

- $Z1$: Relevance: $Cov(X, Z) \neq 0$
  - Z is correlated with X.
- $Z2$: Exogeneity: $Cov(Z, u) = 0$
  - Z is uncorrelated with any unmodelled determinants of Y.
- $Z3$: Exclusion: $Y = \beta_0 + \beta_1 X + \delta Z + u$ has $\delta = 0$
  - Z does not appear in the causal model.

$Z2$ & $Z3$ implies Z cannot have a ***direct*** effect on Y - it can only have an *indirect effect* via X.

$Z1$ implies its effect on X must be non-zero.

Structural Equation

$$Y = \beta_0 + \beta_1 X + u$$

Where OR does not hold, hence,

$$Cov(X, u) \neq 0 \text{ or equivalently } E[Xu] \neq 0$$

First Stage Regression

$$X = \pi_0 + \pi_1 Z + v := X^* + v$$

Where OR holds by construction (population regression), hence,

$$Cov(Z, v) = 0 \text{ or } E[v] = E[Zv] = 0$$

Relevance condition implies $\pi_1 \neq 0$, this does not need to be causally interpreted.

Reduced Form Equation

$$\begin{aligned}Y &= \beta_0 + \beta_1(\pi_0 + \pi_1 Z + v) + u \\ &= (\beta_0 + \beta_1\pi_0) + \beta_1\pi_1 Z + (\beta_1 v + u) \\ &= \gamma_0 + \gamma_1 Z + \epsilon\end{aligned}$$

Where, by the fact that $E[v] = E[Zv] = 0$ and by $Z2$: $Cov(Z, u) = 0$, OR holds, hence,

$E[\epsilon] = E[Z\epsilon] = 0$

And

$\gamma_1 = \beta_1\pi_1$

Estimation Method 1: Indirect Least Squares (ILS)

$$\beta_1 = \frac{\beta_1\pi_1}{\pi_1} = \frac{\gamma_1}{\pi_1} = \frac{Cov(Y, Z)/Var(Z)}{Cov(X, Z)/Var(Z)} = \frac{Cov(Y, Z)}{Cov(X, Z)}$$

$Z1$ is essential here, since $Z1$: $Cov(X, Z) \neq 0$ implies that $\pi_1 \neq 0$ and hence the fraction is well defined.

Sample Counterpart,

$$\hat{\beta}_1 = \frac{\hat{\gamma}_1}{\hat{\pi}_1} = \frac{C\hat{o}v(Y, Z)}{C\hat{o}v(X, Z)}$$

Estimation Method 2: Two-Stage Least Squares (2SLS)

$$X = \pi_0 + \pi_1 Z + v := X^* + v$$

Given that v is uncorrelated with Z (by construction since this a population linear regression) it is also uncorrelated with X^* since X^* is a function of Z.

$$Cov(u, X^*) = cov(u, \pi_0 + \pi_1 Z) = \pi_1 cov(u, Z) = 0$$

Hence X^* gives the part of X that is uncorrelated with u.

$$\begin{aligned}Y &= \beta_0 + \beta_1(X^* + v) + u \\ &= \beta_0 + \beta_1 x^* + (\beta_1 v + u) \\ &= \beta_0 + \beta_1 X^* + \eta\end{aligned}$$

Where,

$$Cov(X^*, \eta) = Cov(X^*, \beta_1 v + u) = 0$$

Hence β_1 can be recovered by a two-stage process,

(1) Population regression of X on Z, giving predicted values X^*.

(2) Population regression of Y on X^*, yielding,

$$\beta_1 = \frac{Cov(Y, X^*)}{Var(X^*)}$$

Sample analogue, $\hat{X} = \hat{\pi}_0 + \pi_1 Z$ (fitted values from regression of X on Z).

$$\hat{\beta}_1 = \frac{C\hat{o}v(Y, \hat{X})}{V\hat{a}r(\hat{X})}$$

2SLS and ILS Coincide in the One Instrument Case

$$\begin{aligned}\beta_1 &= \frac{Cov(Y,X^*)}{Var(X^*)} = \frac{Cov(Y,\pi_0+\pi_1 Z)}{Cov(X^*,X-v)} = \frac{\pi_1 Cov(Y,Z)}{Cov(\pi_0+\pi_1 Z, X-v)} \\ &= \frac{\pi_1 Cov(Y,Z)}{\pi_1 Cov(Z,X)+\pi_1 Cov(Z,v)} \text{ where } Cov(Z,v)=0 \text{ by OR}\end{aligned}$$

$$\beta_1 = \frac{Cov(Y,Z)}{Cov(Z,X)}$$

Multiple Instrument Case

Assumptions

Suppose a variable Z that satisfies the following conditions:

- $Z1$: Relevance: At least one of $\pi_1, ..., \pi_m$ is nonzero.
- $Z2$: Exogeneity: $Cov(Z_I, u) = 0 \;\; \forall I \in \{1, ..., m\}$
  - Z's are uncorrelated with any unmodelled determinants of Y.
- $Z3$: Exclusion: $Z_1, ..., Z_m$ does not appear in the causal model.

Structural Equation

$$Y = \beta_0 + \beta_1 X + \beta_2 W_1 + ... + \beta_{r+1} W_r + u$$

Where X is (potentially) endogenous, and W's are exogenous, hence,

$$E[u] = 0 \;,\; Cov(X, u) \neq 0 \;,\; Cov(W_I, u) = 0 \;\; \forall I \in \{1, ..., r\}$$

First Stage Regression

$$X = \pi_0 + \pi_1 Z_1 + ... + \pi_m Z_m + \pi_{m+1} W_1 + ... + \pi_{m+r} W_r + v$$

OR holds by construction since this a population regression.

Estimation: Two-Stage Least Squares (2SLS)

2SLS is preferable to ILS for multivariate regressions. In fact 2SLS is *needed* when there are more instruments than endogenous variables that they are instruments for.

We start with the first stage regression,

$$X = \pi_0 + \pi_1 Z_1 + ... + \pi_m X_m + \pi_{m+1} W_1 + ... + \pi_{m+r} W_r + v := X^* + v$$

And then substituting this into the structural equation to give,

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \beta_2 W_1 + ... + \beta_{r+1} W_r + u \\ &= \beta_0 + \beta_1 (X^* + v) + \beta_2 W_1 + ... + \beta_{r+1} W_r + u \\ &= \beta_0 + \beta_1 X^* + \beta_2 W_1 + ... + \beta_{r+1} W_r + (\beta_1 v + u) \\ &= \beta_0 + \beta_1 X^* + \beta_2 W_1 + ... + \beta_{r+1} W_r + \epsilon \end{aligned}$$

In order for OR to hold in this equation it must be the case that,

(1) $E[\epsilon] = 0$
  - Which holds because $E[v] = E[u] = 0$

(2) $Cov(W_I, \epsilon) = \beta_1 Cov(W_I, v) + Cov(W_I, u) = 0 \;\; \forall I \in \{1, ..., r\}$
  - $Cov(W_I, u) = 0$ holds by the assumption that the W's are exogenous.
  - $Cov(W_I, v) = 0$ holds by construction since the W's are included in the first stage regression which is a population regression.

(3) $Cov(Z_I, \epsilon) = \beta_1 Cov(Z_I, v) + Cov(Z_I, u) = 0 \ \forall I \in \{1, ..., r\}$

- $Cov(Z_I, u) = 0$ holds by $Z2$.
- $Cov(Z_I, v) = 0$ holds by construction since the first stage regression is a population regression.

(4) No Multicollinearity

- There must be no multicollinearity between the RHS variables in

$$Y = \beta_0 + \beta_1 X^* + \beta_2 W_1 + ... + \beta_{r+1} W_r + \epsilon$$

- Multicollinearity would arise in the case in which $\pi_1 = ... = \pi_m = 0$ as in that case $X*$ is simply a linear combination of the W's.

Sample Analogue,

$$\hat{X} = \hat{\pi}_0 + \hat{\pi}_1 Z_{1i} + ... + \hat{\pi}_m Z_{mi} + \hat{\pi}_{m+1} W_{1i} + ... + \hat{\pi}_{m+r} W_{ri}$$

Then compute $\hat{\beta}_1$ as the coefficient on $\hat{X}$ in a regression of Y on $\hat{X}, W_1, ...W_r$.

Good Instruments?

Natural Experiments

- Great for IVs.
- Exogenous policy changes effect X's.
- Affect X as if randomly.
- Generate Z's that shift X as if randomly.
- Hence uncorrelated with u.

Examples:

(1) Y: Crime rates, X: Incarceration rates, looking at "would higher crime rates lead to higher incarceration rates?"

- Simultaneity problem.
- Prison capacity constraints mean that incarceration rates were kept lower.
- Use Z: lawsuits aimed at reducing prison overcrowding as an IV.

(2) Y: Mortality from heart disease, X: CC (cardiac catheterisation)

- Endogeneity problem: CC treatment depends on doctors assessment of patients overall health, which also impacts mortality (patients overall health is in u).
- Z: distance from patients' home to nearest CC performing hospital minus distance from patients home to any hospital.
  - Relevance: distance plausibly influences likelihood of getting CC.
  - Exogeneity: not correlated with unobserved determinants of morality.
  - Exclusion: expressed as relative distance since actual distance could indeed affect morality.

Simultaneity & Instrumental Variables

Early in the section titled 'Problems with Regression Analysis' we briefly considered the probability of simultaneity causing endogeneity and hence not allowing us to consistently estimate our coefficients.

Problem of Simultaneity: Supply & Demand

Note that in economics we only ever observe the equilibrium values for price and quantity (or wage and hours worked, etc) because we can't go around asking everyone 'how much of the good would you purchase if the price was £x'- we don't live in some counterfactual land.

Hence for supply and demand for milk we have:

$$\begin{aligned} \ln\left(Q^s\right) &= \alpha_0 + \alpha_1 \ln(P) + u \Leftrightarrow q^s = \alpha_0 + \alpha_1 p + u \\ \ln\left(Q^d\right) &= \delta_0 + \delta_1 \ln(P) + v \Leftrightarrow q^d = \delta_0 + \delta_1 p + v \end{aligned}$$

Which, given the equilibrium conditions, $q_i^s = q_i^d$, returns,

$$\begin{aligned} q_i &= a_0 + a_1 p + u \\ q_i &= \delta_0 + \delta_1 p + v \end{aligned}$$

We actually can't tell which is supply and which is demand without other regressors, such as income for the demand equation and weather for the supply equation.

Example: Simultaneity Bias

(1) $y_1 = \alpha y_2 + \beta_1 z_1 + u$

(2) $y_2 = \alpha_2 y_1 + \beta_2 z_2 + v$

$$\begin{aligned} y_2 &= \alpha_2\left(a_1 y_2 + \beta_1 z_1 + u\right) + \beta_2 z_2 + v \\ \left(1 - a_1\alpha_2\right) y_2 &= a_2\beta_1 z_1 + \beta_2 z_2 + v + a_2 u \\ y_2 &= \frac{a_2\beta_1}{(1-a_1a_2)} z_1 + \frac{\beta_2}{(1-a_1a_2)} z_2 + \frac{v + a_2 u}{(1-a_1a_2)} \end{aligned}$$

Now consider whether or not OR will hold in equation (1)

$$\begin{aligned} cov(y_2, u) &= cov\left(\frac{a_2\beta_1}{(1-a_1a_2)} z_1 + \frac{\beta_2}{(1-a_1a_2)} z_2 + \frac{v + a_2 u}{(1-a_1a_2)}, u\right) \\ &= \underbrace{cov\left(\frac{\alpha_2\beta_1}{(1-a_1\alpha_2)} z_1, u\right)}_{=0} + \underbrace{cov\left(\frac{\beta_2}{(1-a_1a_2)} z_2, u\right)}_{=0} + cov\left(\frac{v + a_2 u}{(1-a_1a_2)}, n\right) \\ &= cov\left(\frac{v + a_2 u}{(1-a_1a_2)}, u\right) \neq 0 \end{aligned}$$

Hence OR does not hold.

- Note that we suppose that the covariance of the z_I terms with the error is zero since in equation (1) and (2) we supposed OR held
- In other words we are supposing z_1 *only* affects y_1 and hence is not correlated with u, and vice-versa for z_2 and v.

So regression (1) and (2) will not recover the parameters due to simultaneity, unless $\alpha_2 = 0$ and u and v are uncorrelated, but that is a strong assumption.

Example: Consistently Recovering the Elasticities

We are going to consider the price and quantity of milk, recall that we only observe the equilibrium values of quantity and price since we can't be bothered to go on a mission to ask everyone quantities they would supply or buy in counterfactual situations.

Suppose the regression:

$$\ln\left(Q_i^{\text{milk}}\right) = \delta_0 + \delta_1 \ln\left(P_i^{\text{milk}}\right) + u_i$$
$$q_i = \delta_0 + \delta_1 p_i + u_i$$

We know that δ_1 gives the price elasticity of milk - that is the percentage change in quantity for a 1% change in the price. (This comes from the fact this regression is log-log.)

An OLS regression of this equation suffers a simultaneity problem, since price and quantity are determined by the interaction of demand and supply.

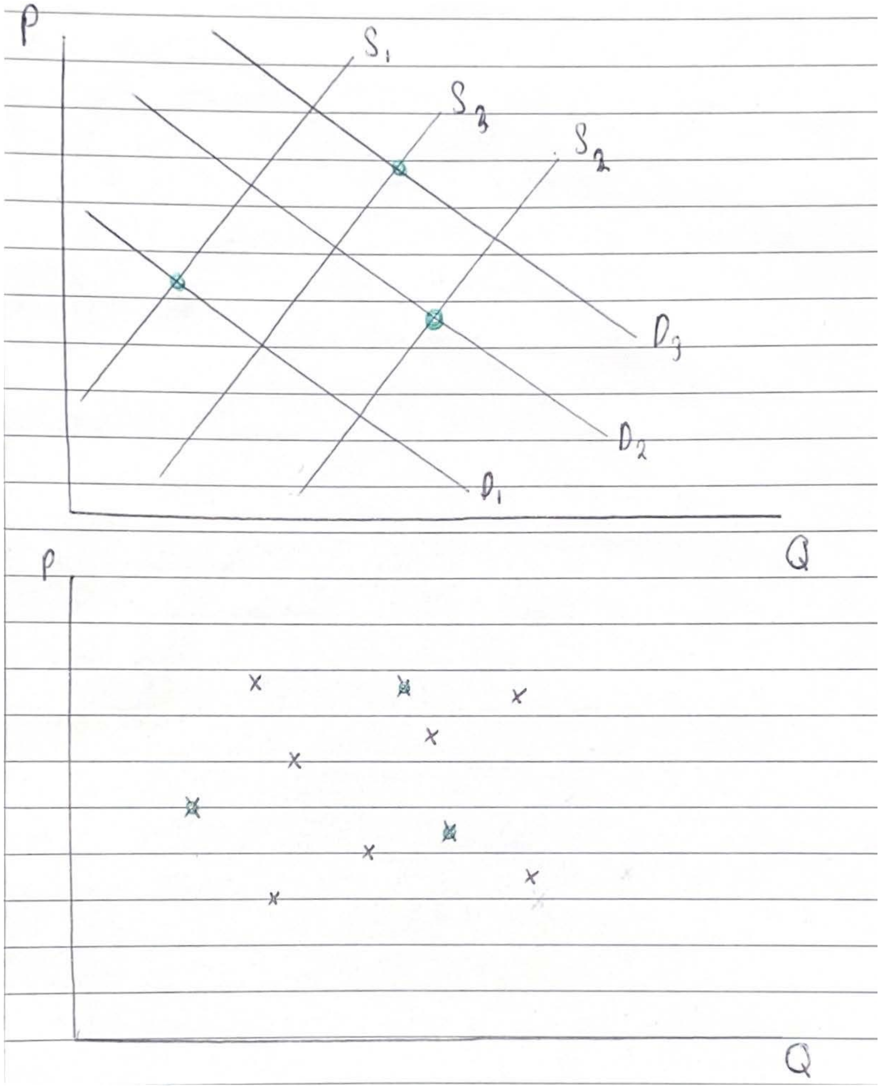

Figure 1: When supply & demand both shift we cannot consistently estimate either

And a regression of these points produced will not recover either the supply or demand curve for us, so we have no idea what is going on. . .

Suppose we want to recover the demand curve.

To consistently recover the demand curve we need to consider an instrument which makes only the supply curve shift since that instrument while generate the points from which we can recover the demand curve and hence its gradient (the elasticity).

2SLS hence estimates the demand curve with the use of an instrument that shifts supply but not demand.

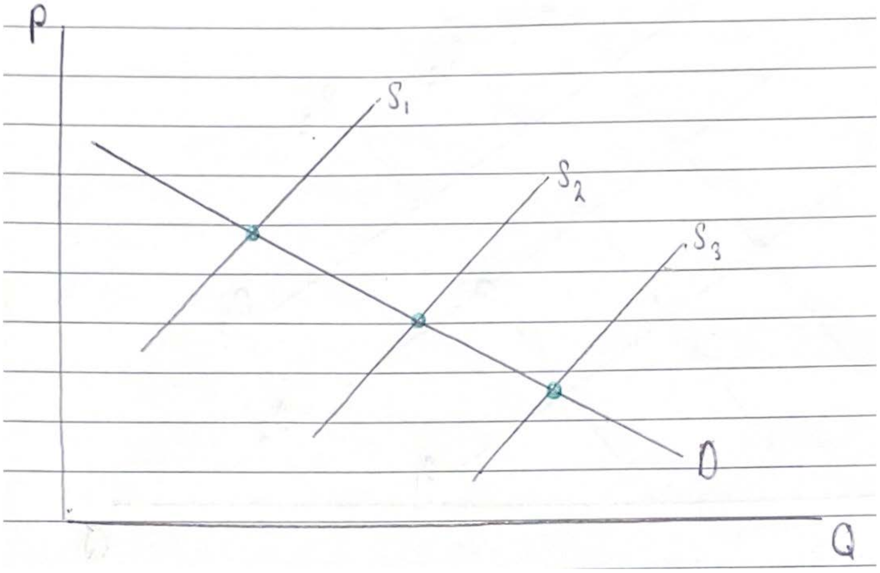

Figure 2: With an instrument that only shifts supply we can consistently estimate demand

Example: Rainfall in dairy producing regions, since rainfall does not affect demand but is plausibly correlated with the amount the cattle can graze and hence how much milk they produce.

Inference in 2SLS Regression

Large Sample Distribution

$$Y = \beta_0 + \beta_1 X + \beta_2 W_1 + \ldots + \beta_{r+1} W_r + u$$

Where X is potentially endogenous, and m instruments $Z_1, \ldots, Z_m$ are available.

Assuming:

- $Z_I \forall I \in \{1, \ldots, m\}$ satisfies $Z1$, $Z2$, $Z3$ (Z's are valid instruments).
- $cov\,(W_I, u) = 0 \quad \forall I \in \{1, \ldots, k\}$ (W's are exogenous).
- Y_i, X_i, Z_i, W_h are iid.
- Large outliers are unlikely.
- There is no perfect multicollinearity among any subset of $(W_1, \ldots, W_r)$ or $(Z_1, \ldots, Z_m)$

Hence,

$$n^{1/2}(\hat{\beta}_1 - \beta_1) \xrightarrow{D} N\left[0, \omega^2_{\beta_{1,N}}\right]$$

Asymptotic variance is complex but can be calculated in R - use heteroskedastic robust option.

Precision

For,

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + u \\ X &= \pi_0 + \pi_1 Z_1 + \ldots + \pi_m Z_m + v \\ X^* &\equiv \pi_0 + \pi_1 Z_1 + \ldots + \pi_m Z_m \end{aligned}$$

Then,

$$\omega^2_{\beta^2_1} = \frac{E\left[(X^* - \mu_{X^*})^2 u^2\right]}{\left(E\left[(X^* - \mu_{X^*})^2\right]\right)^2}$$

Where $\mu_{X^*} = E\left[X^*\right] = E[X]$

For homoskedastic u $\left(E\left[u^2 \mid Z_1, \ldots, Z_m\right] = E\left[u^2\right],\right.$

$$\omega^2_{\beta_{1,N}} = \frac{\text{var}(u)}{\text{var}\left(X^*\right)}$$

Hence the precision of 2SLS improves (the asymptotic variance falls) if

- The 'fit' of the structural equation is better, that is $var(u)$ is lower.
- More of X is explained by $Z_1, \ldots, Z_m$, that is $var\left(X^*\right)$ is higher.
- Using more (valid) instruments: improves 'fit' of first stage, hence $var\left(X^*\right)$ is higher.

2SLS is less efficient than OLS since we are considering,

$$\omega^2_{\beta_{1,IV}} = \frac{\text{var}(u)}{\text{var}\left(X^*\right)} \text{ vs } \omega^2_{\beta_{1,OLS}} = \frac{\text{var}(u)}{\text{var}(X)}$$

And we know that,

$$var\left(\text{X}^*\right) = \text{var(X)} + \text{var(v)} \geq \text{var(X)}$$

Standard errors,

$$s.e.(\hat{\beta}_{1N}) = n^{1/2}\omega_{\beta_{1,IV}}$$

Testing Instrument Validity

Consider the model,

$$Y = \beta_0 + \beta_1 X + \beta_2 W_1 + \ldots + \beta_{r+1} W_r + u$$

Where X is potentially endogenous, W's are exogenous; and there are m possible instruments $Z_1, ..., Z_m$ available.

Suppose that we want to test the validity of our instruments $Z_1, ..., Z_m$, that is we want to test our assumptions $Z1$ and $Z2$ in order to know whether or not our instruments meet them. This is how we might go about that.

Relevance and Weak Instruments

Assumption: Z1

Z1: Relevance: At least one of $\pi_1, \ldots, \pi_m$ is nonzero in the regression

$$X = \pi_0 + \pi_1 Z_1 + \ldots + \pi_m Z_m + \pi_{m+1} W_1 + \ldots + \pi_{m+r} W_r + \nu$$

Testing: F-test

$Z1$ may be tested by the F-test:

$$\begin{aligned} &H_0 : \pi_1 = \ldots = \pi_m = 0 \\ &H_1 : \pi_1 \neq 0 \quad \exists I \in \{1, \ldots, m\} \end{aligned}$$

$$F = \frac{SSR_{rs} - SSR_{un}}{SSR_{un}} \frac{n-k-1}{q} = \frac{(SSR_{rs} - SSR_{un})/q}{SSR_{un}/(n-k-1)} \xrightarrow{d} F_{q,\infty}$$

Where: q is the number of restrictions and k is the number of slope coefficients in unrestricted model

Problem: Weak Instruments

Showing that the instruments aren't irrelevant doesn't help us enough...

If $Z1$ (relevance) holds, but only by a small amount, does not justify that inferences on β_1 based on 2SLS estimates and standard errors are necessarily reliable.

Why?

- For one Z and one W
$$\omega^2_{\beta_{1,IV}} = \frac{\operatorname{var}(u)}{\operatorname{var}(X^*)} = \frac{\operatorname{var}(u)}{\pi_1^2 \operatorname{var}(Z)}$$
- In this simplified setting if π_1 is very close to zero then the asymptotic variance $\omega^2_{\beta_{1,IV}}$ will be very large and $\hat{\beta}_1$ will not be asymptotically normal.
- This is known as having a **weak instrument**.

Solution

- No longer compare F statistic to critical values drawn from $F_{m,\infty}$ distribution, but rather use larger critical values.
- Often use 10 as a rule of thumb.

Instrument Exogeneity

Assumption: Z2

Instrument exogeneity is implied by $Z2$ and requires,

$$cov\left(Z_I, u\right) = 0 \quad \forall \in \{1, \ldots, m\}$$

Testing

We might want to test this by considering if

$$c\hat{o}v\left(Z_I, \hat{u}\right) = 0 \quad I \in \{1, \ldots, m\}$$

Where $\hat{u} = Y - \hat{\beta}_0 - \hat{\beta}_1 X - \hat{\beta}_2 W_1 - ... - \hat{\beta}_{r+1} W_r$

Of course if it is the case that $c\hat{o}v\left(Z_I, \hat{u}\right) = 0 \quad \forall I \in \{1, \ldots, m\}$ then our $Z2$ assumption is valid.

Problem: Single Instrument Case

The problem is that for a single instrument the restriction that $cov(Z, u) = 0$ is wholly used up in constructing the 2SLS estimates.

Hence we cannot test for exogeneity of a single instrument.

(I) Explanation:

$$c\hat{o}v(Z, \hat{u}) = c\hat{o}v\left(Z, Y - \hat{\beta}_0 - \hat{\beta}_1 X\right) = c\hat{o}v(Z, Y) - \hat{\beta}_1 c\hat{o}v(Z, X)$$

- But recall that,

$$\hat{\beta}_1 = \frac{c\hat{o}v(Z, Y)}{c\hat{o}v(Z, X)}$$

- So it is the case that

$$c\hat{o}v(Z, \hat{u}) = c\hat{o}v(Z, Y) - \frac{c\hat{o}v(Z, Y)}{c\hat{o}v(Z, X)} c\hat{o}v(Z, X) = 0$$

  What this means is that it is *always the case* in the single instrument IV regression that $c\hat{o}v(Z, \hat{u}) = 0$ by construction.

This implies we cannot test whether or not $c\hat{o}v(Z, \hat{u}) = 0$ and so we can't test the exogeneity condition with only one instrument.

Solution: Multiple Instrument Case

We can only test instrument exogeneity when we have multiple instruments (more than one Z)

(I) Explanation (1):

$$Y = \beta_0 + \beta_1 X + u$$
$$\hat{X} = \hat{\pi}_0 + \hat{\pi}_1 Z_1 + \hat{\pi}_2 Z_2$$

- By the preceeding argument it must be the case that $c\hat{o}v(\hat{X}, \hat{u}) = 0$, but it doesn't necessarily follow that $c\hat{o}v\left(Z_1, \hat{u}\right) = c\hat{o}v\left(Z_2, \hat{u}\right) = 0$

(II) Explanation (2):

- If both IV's are valid then:
$$\frac{\text{cov}\,(Y,Z_1)}{\text{cov}\,(X,Z_1)} = \beta_1 = \frac{\text{cov}\,(Y,Z_2)}{\text{cov}\,(X,Z_2)}$$
- Using Z_1 and Z_2 as separate instruments then:
$$\hat{\beta}_{1\mid Z_1} = \frac{\hat{cov}\,(Y_1 Z_1)}{\hat{cov}\,(X,Z_1)}\ ,\ \hat{\beta}_{1\mid Z_2} = \frac{\hat{cov}\,(Y,Z_2)}{\hat{cov}\,(X,Z_2)}$$
- Under exogeneity both of these estimates should be close to one another.

Testing: F-test

Assume homoskedasticity : $E\left[u^2 \mid Z_1,\ldots,Z_m\right] = E\left[u^2\right]$

1. Compute 2SLS residuals:
$$\hat{u} = Y - \hat{\beta}_0 - \hat{\beta}_1 X - \hat{\beta}_2 W_1 - \ldots - \hat{\beta}_{r+1} W_r$$
2. Conduct (homoskedastic) F-test
$$\begin{aligned} &H_0 : \delta_1 = \ldots = \delta_m = 0 \\ &H_1 : \delta_I \neq 0 \quad \exists I \in \{1,\ldots,m\} \end{aligned}$$
$$\begin{aligned} &UN : \hat{u} = \delta_0 + \delta_1 Z_1 + \ldots + \delta_m Z_m + \delta_{m+1} W_1 + \ldots + \delta_{m+r} W_r + \eta \\ &RS : \hat{u} = \delta_0 + \delta_{m+1} W_1 + \ldots + \delta_{m+r} W_r + \eta \end{aligned}$$

Use the adjusted F-statistic:
$$F^* = \frac{m}{m-1}F$$
Which basically just changes our old F- stat to have one less degree of freedom

And compare to critical values from $F_{m-1,\infty}$, because we already said this won't work for $m = 1$ – one restriction is used up in estimating by 2SLS hence we have one less degree of freedom.

Hetero-robust version has distribution:
$$\mathcal{X}^2_{m-1}$$

Randomised Control Trials: Imperfect Compliance

Treatment Status vs Offer to Treat

D is the treatment status, that is whether the treatment was actually received.

Z is the offer to treat / treatment assignment.

Perfect compliance means that $D_i = Z_i \quad \forall i$; imperfect compliance means that $D_i \neq Z_i \quad \exists i$.

We can always randomly assign Z, but we cannot always randomly assign D

- **Example:** We can always offer some people randomly discounted University fees, but we can't assign that some people attend university and others do not - ethically that would not be okay.

Estimated Treatment Effect (2SLS) and Intention to Treat (ITT)

	(1)	(2)	(3)	(4)
Dep. var	Y	Y	D	Y
Method	OLS	2SLS	OLS	OLS
D	0.087	0.145	0.786	
	(0.044)	(0.060)	(0.043)	
Z				0.108
				(0.041)

Here (2) gives estimated treatment effect (2SLS), that is,

- How Y changes with D,
- How the variable we are interested in impacts the dependent variable.

And here (4) estimates the ITT: Intention to treat (OLS)

- How Y changes with Z,
- How the inducement impacts the dependent variable.

	(1)	(2)	(3)	(4)
Dep. var	lwage	luage	educ	lwage
Method	OLS	2SLS	OLS	OLS
educ	0.079	0.145		
	(0.002)	(0.063)		
Z			0.436	0.064
			(0.071)	(0.026)

So (1) gives inconsistent OLS estimate,

But (2) gives consistent 2SLS estimate (Treatment effect),

And (3) first stage regression,

While (4) reduced form regression (ITT),

And notice that,

$$\frac{0.064}{0.436} = 0.145$$

Hence either method recovers the coefficient in this simple regression mode. That is

$$\frac{\text{ITT}}{\text{First Stage}} = \text{ILS or 2SLS}$$

When Random Assignment fails

Fails due to:

(1) Non-compliance with the treatment protocol

- Example: Pushy parents get their children into the class with the teacher they think is best.

(2) Impossible to realise due to costs/ethics

- Example: Completing a degree, attending private vs state school...
- Can't randomly assign students to not go to uni/drop out as a control group to see if a degree helps wage prospects.

Solution: Randomly assign an inducement

Assign the payments independently of student's characteristics.

If the **inducement affects uptake of the treatment** then it is an **instrumental variable.**

- E.g. Offer a sum of money for people to complete a degree.

If this attracts people to do degrees then use the inducement as the IV.

Z can be randomly assigned

- E.g. Award scholarships to university by a lottery.

Use Z and the instrument for D

Heterogeneous Causal Effects

$$\begin{aligned} Y_i &= \beta_0 + \beta_{1i} X_i + u_i \\ X_i &= \pi_0 + \pi_1 Z_i + v_i \\ \hat{\beta}_{IV} &= \frac{c\hat{o}v(Y,Z)}{c\hat{o}v(X,Z)} \xrightarrow{p} \frac{cov(Y,Z)}{cov(X,Z)} = \beta_{IV} \\ \beta_{IV} &= E\left[\beta_{1i} \cdot \frac{\pi_{1i}}{E\left[\pi_{1i}\right]}\right] = \text{ LATE} \end{aligned}$$

2SLS estimates the average treatment effect of those who respond most to the instrument (the offer of treatment).

It's a weighted average of the underlying heterogeneous causal effects termed the **local average treatment effect.**

Time Series: Stationarity

Strict Stationarity

The time series $\{Y_t, t \in \mathbb{Z}\}$ is strictly stationary if the joint distributions $(Y_t, Y_{t+1}, \ldots, Y_{t+k}) \underset{D}{=} (Y_s, Y_{s+1}, \ldots, Y_{s+k})$ for all t, s and k

Weak Stationarity

The time series $\{Y_t, t \in \mathbb{Z}\}$ is weakly stationary if:

(i) $E[Y_t] = m$ for all t

(ii) $Var(Y_t) = \sigma^2 < \infty$ for all t.

(iii) $Cov(Y_t, Y_s) = Cov(Y_{t+h}, Y_{s+h})$ for all $t, s, h \in \mathbb{Z}$.

(iii') $Cov(Y_t, Y_{t-h}) = \gamma_h$ for all h (this is equivalent to (iii)).

Descriptive Statistics

Given that $\{Y_t\}_{t=1}^{T} = \{Y_t\}$ is stationary we have,

Sample Mean

$$\bar{Y}_T = \frac{1}{T}\sum_{t=1}^{T} Y_t \xrightarrow{P} \mu = E\left[Y_t\right]$$

Sample Variance

$$\hat{\gamma}_0 = V\hat{a}r\left(Y_t\right) = \frac{1}{T}\sum_{t=1}^{T}\left(Y_t - \bar{Y}_T\right)^2$$

Notice that the variance is equivalent to covariance ($Cov(Y_t, Y_{t-h})$) at $h = 0$, hence it is the same as γ_0.

Sample Covariance

$$\hat{\gamma}_h = c\hat{o}v\left(Y_t, Y_{t-h}\right) = \frac{1}{T}\sum_{t=h+1}^{T}\left(Y_t - \bar{Y}_T\right)^2\left(Y_{t-h} - \bar{Y}_T\right)^2$$

hth autocorrelation function (ACF);

$$\rho_h = \frac{Cov\left(Y_t, Y_{t-h}\right)}{\left[Var\left(Y_t\right) Var\left(Y_{t-h}\right)\right]^{1/2}} = \frac{Cov\left(Y_t, Y_{t-h}\right)}{Var\left(Y_t\right)}$$

With the sample analogue,

$$\hat{p}_h = \frac{c\hat{o}v\left(Y_t, Y_{t-h}\right)}{v\hat{a}r\left(Y_t\right)} = \frac{\hat{\gamma}_h}{\hat{Y}_0}$$

Persistence

We measure persistence using the ACF, which measures the extent of correlation between Y_t and Y_{t-h} as h varies.

We only need to consider positive h values because, $\gamma_h = Cov\left(Y_t, Y_{t-h}\right) = Cov\left(Y_{t+h}, Y_t\right) = Cov\left(Y_t, Y_{t+h}\right) = \gamma_{-h}$ which implies it is the case that $\rho_h = \rho_{-h}$. Note also that $\rho_0 = \operatorname{cov}\left(Y_1, Y_1\right) / \operatorname{var}\left(Y_1\right) = 1$

Persistence is,

- The speed at which {Y(t)} reverts to its mean
- The extent of serial correlation in the time series

ρ_h decays more gradually as h increases for more persistent series.

Weakly stationary models tend to be only weakly persistent (correlated at short lags, but have tendency to revert to the mean)

Autoregressive Models

AR(1) & AR(p) Model

Models

$$\text{AR}(1) : Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$$

$$\begin{aligned}\text{AR(p)} : Y_t &= \beta_0 + \beta_1 Y_{t-1} + \ldots + \beta_p Y_{t-p} + u_i \\ Y_t &= \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + u_t\end{aligned}$$

For $t \in \{1, 2, \ldots, T\}$, with Y_0 as the initial value, which is also a random variable and where $\{u_t\}$ is the driving innovation or shock sequence and is assumed to be stationary and not forecastable based on past values of Y,

$$0 = E\left[u_t \mid Y_{t-1}, Y_{t-2}, \ldots\right] = E\left[u_t \mid y_{t-1}\right]$$

Or we can assume $\{u_t\}$ is iid, hence

$$E\left[u_r\right] = 0 \ , \ E\left[u_t^2\right] = \sigma_u^2$$

Which implies the conditional expectation above.

This implies that $\{u_t\}$ is serially uncorrelated

$$Cov\left(u_t, u_{t-h}\right) = 0 \quad \forall h \neq 0$$

Stationarity

Stationarity of AR(1) requires,

$$\begin{aligned}&\beta_1 \in (-1, 1) \\ &Y_0 \text{ is such that } E[Y_0] = \frac{\beta_0}{1-\beta_1} \ , \ Var(Y_0) = \frac{\sigma_u^2}{1-\beta_1^2}\end{aligned}$$

Proof:

- Assume that $\{Y_t\}$ is stationary, hence

$$\begin{aligned}\mu_Y &= E\left[Y_t\right] = E\left[Y_{t-1}\right] \\ \sigma_Y^2 &= \text{var}\left(Y_t\right) = \text{var}\left(Y_{t-1}\right) \\ Y_h &= \text{cov}\left(Y_t, Y_{t-h}\right)\end{aligned}$$

- Hence it must be the case that,
  - Expectation

$$\begin{aligned}E[Y_t] &= \beta_0 + \beta_1 E[Y_{t-1}] + \underset{=0}{E[u_t]} \Rightarrow \mu_Y = \beta_0 + \beta_1 \mu_Y \\ E[Y_0] &= \mu_Y = \frac{\beta_0}{1-\beta_1}\end{aligned}$$

  - Variance

$$\begin{aligned}Var(\gamma_t) &= \beta_1^2 Var(\gamma_{t-1}) + Var(u_i) + 2\beta_1 Cov(\gamma_{t-1}, u_i) \\ &= \beta_1^2 Var(Y_{t-1}) + Var(u_i) \\ \sigma_Y^2 &= \beta_1^2 \sigma_Y^2 + \sigma_u^2 \\ Var(Y_0) &= \sigma_Y^2 = \frac{\sigma_u^2}{1-\beta_1^2} \Rightarrow |\beta_1| \geq 1\end{aligned}$$

- Autocovariance

$$\begin{aligned}
\operatorname{cov}(Y_t, Y_{t+h}) &= \operatorname{cov}(Y_t, \beta_0 + \beta_1 Y_{t+h-1} + u_{t+h}) = \beta_1 \operatorname{cov}(Y_t, Y_{t+k-1}) \\
&= \beta_1 \operatorname{cov}(Y_t, Y_{t+h-1}) \\
&= \beta_1^2 \operatorname{cov}(Y_t, Y_{t+h-2}) \\
&= {\beta_1}^h \operatorname{cov}(Y_t, Y_t) = \beta_1^h \sigma_Y^2
\end{aligned}$$

Stationarity of AR(p) is a much more involved proof, but one important requirements is that,

$$\sum_{i=1}^{p} \beta_i < 1$$

Forecasting with Autoregressive Models

General Problem

We observe $t = 1, \ldots, T$ and want to forecast $T + h$.

Optimal forecast is mean-squared (forecast) error minimising (MSFE-minimising forecast), hence we need a function $m^* (y_{t-1})$ that solves

$$\underset{m(.)}{min}\, E\left[Y_t - m\left(y_{t-1}\right)\right]^2$$

Which is the conditional expectation - if you don't know why this is the case read over the early proof by clicking here.

Importantly for this section I will be using this shorthand $Y_{T+h\mid T}$ for the conditional expectation of Y_{T+h} given that we know y_T, where $y_T = \{y_T, y_{T-1}, ..., y_1\}$. That is,

$$Y_{T+h\mid T} = E\left[Y_{T+h} \mid y_T\right]$$

AR(p) Case

Assume $\{Y_t\}$ follows,

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + u_i \ \ , \ \ E[u_t \mid y_{t-1}] = 0$$

Hence we know that

$$\begin{aligned}E[Y_t \mid y_{t-1}] &= \beta_0 + \sum_{i=1}^{p} \beta_i \underbrace{E[Y_{t-i} \mid y_{t-1}]}_{=Y_{t-i} \text{ for } i\geq 1} + \underbrace{E[u_t \mid y_{t-1}]}_{=0} \\ &= \beta_0 + \sum_{i=1}^{n} \beta_i Y_{t-i}\end{aligned}$$

Then the optimal 1-step ahead forecast is,

$$\begin{aligned}Y_{T+1\mid T} = E\left[Y_{T+1} \mid y_T\right] &= \beta_0 + \sum_{i=1}^{p} \beta_i Y_{T+1-i} \\ &= \beta_0 + \beta_1 Y_T + ... + \beta_p Y_{T+1-p}\end{aligned}$$

The optimal h-step ahead forecast is,

$$\begin{aligned}Y_{T+h\mid T} = E\left[Y_{T+h} \mid y_T\right] &= \beta_0 + \sum_{i=1}^{p} \beta_i E\left[Y_{T+h-i} \mid Y_T\right] \\ &= \beta_0 + \sum_{i=1}^{p} \beta_i Y_{T+h-i\mid T} \\ &= \beta_0 + \beta_1 Y_{T-1+h} + ... + \beta_p Y_{T-p+h}\end{aligned}$$

Estimated Counterpart & Recursion

There are two important things to note here,

(1) These forecasts are infeasible

- While they are indeed optimal (min MSFE) they are infeasible since we don't know what $\beta_0, ..., \beta_p$ are!

  - Hence we have to use an estimated counterpart to make our forecasts,

$$\hat{Y}_{T+h \mid T} = \hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_j \hat{Y}_{T+h-i \mid T}$$

  - We can do this by OLS since our assumption of $E[u_t \mid y_{t-1}] = 0$ implies that $E[u_t] = 0$ and $Cov(u_t, Y_{t-i}) = 0 \quad \forall i \in \{1, \ldots, p\}$ and hence OR is satisfied.

(2) Need for recursion in the h-step ahead forecast

  - Notice in the h-step ahead forecast, say for $p = 1$ and $h = 2$ we have

$$Y_{T+2 \mid T} = \beta_0 + \beta_1 Y_{T+1 \mid T}$$

  - Notice that as well as not knowing what β_0 and β_1 are, we also don't know what $Y_{T+1 \mid T}$ is as it falls outside of our sample. Hence we will need to use recursion to find future values: Start with your final observation Y_T; Sub that in to find $Y_{T+1 \mid T}$; Repeat for $Y_{T+2 \mid T}$.

Out of Sample Forecast Errors

These errors are for an AR(p) model, with a sample from $\{1, ..., T\}$, and for a one-step ahead forecast for simplicity. This is because in the one step ahead forecast the independent variables in our model: Y_{T+1-i}'s; are all within our dataset for $i = \{1, ..., p\}$ because $Y_{T+1-1} = Y_T$, etc. This means we don't need to use recursion here.

The error from the infeasible forecast is,

$$\begin{aligned} e_{T+1 \mid T} = \left(Y_{T+1} - Y_{T+1 \mid T}\right) &= \left[\beta_0 + \sum_{i=1}^{p} \beta_i Y_{T+1-i} + u_{T+1}\right] - \left[\beta_0 + \sum_{i=1}^{p} \beta_i Y_{T+1-i}\right] \\ &= u_{T+1} \end{aligned}$$

- In case it isn't obvious what is happening here: the error in our forecast - that is the **difference** between what Y_{T+1} actually is, and what we **expected** Y_{T+1} to be: $Y_{T+1 \mid T}$ - in the case in which **we know what β_1 and all β_i are** is just u_{T+1}. That is **the unknown & unforecastable error term that takes us from Y_T to Y_{T+1}**.

The error from the feasible forecast is,

$$\begin{aligned} \hat{e}_{T+11T} = Y_{T+1} - \hat{Y}_{T+1 \mid T} &= \left(\beta_0 + \sum_{i=1}^{p} \beta_i Y_{T+1-i} + u_{T+1}\right) - \left(\hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_i Y_{T+1-i \mid T}\right) \\ &= u_{T+1} + \left\{(\beta_0 - \hat{\beta}_0) + \sum_{i=1}^{p} (\beta_i - \hat{\beta}_i) Y_{T+1-i}\right\} \end{aligned}$$

- Again to explain: Now we have two components to our forecast error.
  - (i) The first is the error we saw in the infeasible forecast case: u_{T+1}. In the feasible forecast case we still don't know what the **the unknown & unforecastable error term that takes us from Y_T to Y_{T+1}** is, so that error is still part of our forecast error.
  - (ii) The second is a new error though. This is estimation error from our OLS regression we ran to estimate what we think β_0 and all the β_i's are. So this is **the difference between what we thought these values were: $\hat{\beta}_0, \hat{\beta}_1, ...$** and what they **actually turned out to be: $\beta_0, \beta_1, ...$**.

The two components are orthogonal (uncorrelated) since $E\left[e_{T+1} \mid y_T\right] = E\left[u_{T+1} \mid y_T\right] = 0$

Forecast performance: MSFE

We really consider the forecast performance by considering the Mean Squared Forecast Error, that is the expectation of the square of the error we found above.

In the infeasible case,

$$\text{MSFE}(Y_{T+1 \mid T}) = E[e^2_{T+1 \mid T}] = E[u^2_{T+1}] = \sigma^2_u$$

And in the feasible case,

$$\begin{aligned}\widehat{\text{MSFE}}(\hat{Y}_{T+1 \mid T}) = E[\hat{e}^2_{T+1 \mid T}] &= E[e^2_{T+11T}] + E[Y_{T+11T} - \hat{Y}_{T+1 \mid T}]^2 \\ &= \sigma^2_u + E\left[Y_{T+11T} - \hat{Y}_{T+11T}\right]^2 \\ &= \sigma^2_u + E[(\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)Y_T]^2\end{aligned}$$

Notice that $\hat{\sigma}^2_u$ isn't a good estimate of $\text{MSFE}(\hat{Y}_{T+1 \mid T})$, it will always underestimate it since if we just used $\hat{\sigma}^2_u$ we miss out the second component of the error - the OLS estimation error.

Estimating MSFE

First we will need to obtain a sequence of 'pseudo' out of sample forecast errors - recall that we can't recover $\hat{e}_{T+1 \mid T} = Y_{T+1} - \hat{Y}_{T+1 \mid T}$ since $Y(T+1)$ is outside of our sample.

So instead we'll solve for some $s < T$ and use that! $\hat{e}_{s+1 \mid T} = Y_{s+1} - \widehat{Y}_{s+1 \mid T}$ where $Y(s+1)$ is within our sample!

If we suppose an AR(p),

-Compute $\hat{e}_{s+1 \mid T} = Y_{x+1} - \hat{Y}_{s+1 \mid T}$

- By estimating a model in the subsample $\{Y_T\}^s_{t=1}$

$$\hat{Y}_{s+1 \mid T} = \widehat{\beta}_{0,s} + \sum_{i=1}^{p} \widehat{\beta}_{i,s} \widehat{Y}_{s+1-i}$$

  - $\hat{e}_{s+1 \mid T} = Y_{s+1} - \hat{Y}_{s+1 \mid T}$ is recoverable since Y_{s+1} is observed.

- Repeat this for the final P periods (for $s \in \{T-P, \ldots, T-1\}$), so we get $\left\{\hat{e}_{s+1 \mid s}\right\}^{T-1}_{s=T-P}$

Finally estimate MSFE:

For $h = 1$

$$\widehat{\text{MSFE}}(\hat{Y}_{T+1 \mid T}) = \frac{1}{P} \sum_{s=T-P}^{T-1} \hat{e}^2_{s+1 \mid T}$$

For any h

$$\widehat{\text{MSFE}}(\hat{Y}_{T+h \mid T}) = \frac{1}{P} \sum_{s=T-P-(h-1)}^{T-h} \hat{e}^2_{s+1 \mid T}$$

Model Selection

How many lags?

Big p means a more flexible model, with less bias.

Small p means a lower variance.

Don't use R^2, SER, or $\bar{R}^2$: The first doesn't penalise larger models at all, the other two don't penalise additional lags enough.

Methods of selecting models

(1) Directly compare $\widehat{\text{MSFE}}(\bar{Y})$

- Problem: may be unreliable if small P is used to estimate the MSFE$(\bar{Y})$ - where P is the sample size used, or the number of periods over which we are estimating the MSFE (note the difference between p and P here).

(2) Stepwise testing down procedure

- Start with some p lags.
  (i) Peform a t-test that $H_0 \ : \ \beta_p = 0$
  (ii) If we accept $H_0 \ : \ \beta_p = 0$ then test $H_0 \ : \ \beta_{p-1}$ and repeat until we reject the null.
- Problem: $\hat{\beta}_i$ could be significant by chance.

(3) Information Criteria

- $AIC_m = log\left(\frac{SSR_m}{T}\right) + m\frac{2}{T}$
- $BIC_m = log\left(\frac{SSR_m}{T}\right) + m\frac{log(T)}{T}$
  - m : number of model parameters (AR(p) has $m = p - 1$).
  - T : length of period, i.e. $1, ..., T$.
- Differences
  - BIC penalises larger models more than AIC,
  - BIC is consistent (it chooses the correct p for an AR(p) with high probability),
  - AIC is conservative (chooses a p slightly larger than correct p).

ARDL(p,q) Model

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{i=1}^{q} \delta_i X_{t-i} + u_i \text{ where } E\left[u_t, y_{t-1}, x_{t-1}\right] = 0 \text{ and } \{Y, X_t\} \text{ are jointly stationary}$$

Standard inference holds if Y and X are stationary and weakly dependent.

MSFE-minimising forecast of Y_{T+1} is,

$$\begin{aligned} Y_{T+1 \mid T} &= E\left[Y_{T+1} \mid y_T, x_T\right] \\ &= E\Big[\beta_0 + \sum_{i=1}^{p} \beta_i Y_{T+1-i} + \sum_{i=1}^{p} \delta_i X_{T+1-i} + u_{T+1} \Bigm| y_T, x_T\Big] \\ &= \beta_0 + \sum_{i=1}^{p} \beta_i Y_{T+1-i} + \sum_{i=1}^{p} \delta_i X_{T+1-i} \end{aligned}$$

Estimate by plugging in OLS estimates:

$$\hat{Y}_{T+1 \mid T} = \hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_i Y_{T+1-i} + \sum_{i=1}^{p} \hat{\delta}_i X_{T+1-i}$$

Get longer run forecasts by recursion

$$\begin{aligned} Y_{T+2 \mid T} &= E\Big[\beta_0 + \sum_{i=1}^{p} \beta_i Y_{T+2-i} + \sum_{i=1}^{p} \delta_i X_{T+2-i} + u_{T+2} \mid y_T, x_T\Big] \\ &= \beta_0 + \sum_{i=1}^{p} \beta_i E\left[Y_{T+2-i} \mid y_T, x_T\right] + \sum_{i=1}^{p} \delta_i E\left[X_{T+2-i} \mid y_T, x_T\right] \\ &= \beta_0 + \sum_{i=1}^{p} \beta_i Y_{T+2-i \mid T} + \sum_{i=1}^{p} \delta_i X_{T+2-i \mid T} \end{aligned}$$

For $p = q = 1$ this would be,

$$\begin{aligned} Y_{T+2 \mid T} &= E\left[\beta_0 + \beta_1 Y_{T+1} + \delta_1 X_{T+1} + u_{T+2} \mid y_T, x_T\right] \\ &= \beta_0 + \beta_1 E\left[Y_{T+1} \mid y_T, x_T\right] + \delta_1 E\left[X_{T+1} \mid y_T, x_T\right] \\ &= \beta_0 + \beta_1 Y_{T+1 \mid T} + \delta_1 X_{T+1 \mid T} \end{aligned}$$

- In order to complete this we'd need to forecast the dependent variable X for $T + 1$ from the model,

$$\hat{X}_t = \hat{\gamma}_0 + \sum_{i=1}^{p'} \hat{\gamma}_i X_{t-i} + \sum_{i=1}^{q'} \hat{\theta}_i Y_{t-i} + v_t \text{ where } E\left[v_t \mid y_{t-1}, x_{t-1}\right] = 0$$

- And of course we would need to forecast Y for $T + 1$ which, as shown above, is,

$$\hat{Y}_{T+1 \mid T} = \hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_i Y_{T+1-i} + \sum_{i=1}^{p} \hat{\delta}_i X_{T+1-i}$$

Granger Causality

For Granger causality we are asking if introducing $\{X_t\}$ help forecast $\{Y_t\}$, that is does switching from an AR(p) to and ARDL(p,q) help us forecast Y?

$\{X_t\}$ **does not Granger cause** $\{Y_t\}$ if lags of $\{X_t\}$ carry no useful information about $\{Y_t\}$ in addition to that carried by its own lags, that is

$$E\Big[\big\{Y_{T+1} - E\left[Y_{T+1} \mid y_T, x_T\right]\big\}^2\Big] = E\Big[\big\{Y_{T+1} - E\left[Y_{T+1} \mid y_T\right]\big\}^2\Big]$$

or

$$\widehat{\text{MSFE}}(Y_{T+1 \mid y,x}) = \widehat{\text{MSFE}}(Y_{T+1 \mid y})$$

$\{X_t\}$ **does Granger cause** $\{Y_t\}$ if lags of $\{X_t\}$ carry useful information about $\{Y_t\}$ in addition to that carried by its own lags, that is

$$E\Big[\big\{Y_{T+1} - E\left[Y_{T+1} \mid y_T, x_T\right]\big\}^2\Big] < E\Big[\big\{Y_{T+1} - E\left[Y_{T+1} \mid y_T\right]\big\}^2\Big]$$

or

$$\widehat{\text{MSFE}}(Y_{T+1 \mid y,x}) < \widehat{\text{MSFE}}(Y_{T+1 \mid y})$$

Testing for Granger Causality:

We must test an ARDL(p,q) where $p = q$, otherwise extra predictability could just come from having more lags.

(1) Choose $p = q$ by sequential testing or by using AIC/BIC.

(2) Perform an F-test,

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{i=1}^{q} \delta X_{t-i} + u_i \text{ where } E\left[u_t \mid y_{t-1}, x_{t-1}\right] = 0 \text{ and } \{Y_t, X_t\} \; : \; \text{jointly stationary}$$

$$H_0 : \delta_1 = \delta_2 = \ldots = \delta_p = 0$$

(3) Rejecting null implies that $\{X_t\}$ Granger causes $\{Y_t\}$.

Nonstationary Times Series

Modelling and forecasting for non-stationary time series

Differencing Non-Stationary to Stationary

We can account for non-stationarity by transforming non-stationary series to stationary ones via,

(1) Differencing

- First difference: $\Delta Y_t = Y_t - Y_{t-1}$
- Seasonal difference: $\Delta_4 Y_t = Y_t - Y_{t-4}$

(2) Logarithms and Growth rates

- Difference of logs:

$$\begin{aligned}\Delta log(Y_t) &= log(Y_t) - log(Y_{t-1}) \\ &= log(\frac{Y_t}{Y_{t-1}}) = log(\frac{\Delta Y_t + Y_{t-1}}{Y_{t-1}}) = log(1 + \frac{\Delta Y_t}{Y_{t-1}}) \approx \frac{\Delta Y_t}{Y_{t-1}}\end{aligned}$$

- Difference of logs is better than percentage changes since logs are symmetric, whereas percentage changes are not,
- If something goes up by 1% from t to $t+1$ (call this 'in $t+1'$), then goes down by 1% in $t+2$ we do not get back to the original value at t
- If a log is 0.5 in $t+1$ and -0.5 in $t+2$ then we get back to the original value at t.
- Change in log ×100 is % growth rate though.
- Annualised % growth rate: $400 \times \Delta \log (Y_i)$.

(3) Deterministic Detrending

$$Y_t - \delta_0 - \delta_1 t$$

Breaks and Parameter Instability

We might also be able to accommodate non-stationarity by checking if these stationarity holds during certain epochs. That is by checking if there are breakpoints at which the series becomes non-stationary.

For this we will consider the model,

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t \ , \ E\left[u_i, y_{t-1}\right] = 0 \ , \ E\left[u_t^2\right] = \sigma_u^2 \ , \ |\beta_1| < 1$$

Known Break: Breakpoint Dummies

Suppose we know that the break is at period τ

We could then specify AR(1) model with breakpoint dummies

$$\gamma_t = \beta_0 + \beta_1 Y_{t-1} + \gamma_0 D_t(\tau) + \gamma_1 D_t(\tau) Y_{t-1} + u_t$$

$$D_t(\tau) = \begin{cases} 1 \text{ if } t \leq \tau \\ 0 \text{ if } t > \tau \end{cases}$$

The implication of this is that:

- Intercept: $\beta_0 + \gamma_0$ until τ then β_0.
- Coefficient: $\beta_1 + \gamma_1$ until τ then β_1.

Chow Test

On the same AR(1) model above we would test

$$H_0 : \gamma_0 = \gamma_1 = 0$$

With an F-test $F_{2,\infty}$ (2 because there are 2 restrictions)

Accepting null implies no break.

Unknown Break

Suppose we do not know where a break could be. How might we go about testing in this case?

QLR Test

Allow for breaks at $\tau \in \{\tau_0, \tau_0 + 1, \ldots, \tau_1\}$ for $\tau_0 \simeq \pi T, \tau_1 \simeq (1 - \pi)T$

For each τ compute the Chow test statistic $F(\tau)$

QLR statistic is the largest of all of these Chow test statistics:

$$QLR = \max \{F(\tau_0), F(\tau_0 + 1), \ldots, F(\tau_1)\}$$

Tend to choose $\pi = 15\%$

Critical values depend on q (the number of restrictions)

$$\hat{\tau} = argmax \; F(\tau)$$

Unit Roots & Stochastic Trends

Deterministic trends: has a determined trend, such as $y_t = ct + \varepsilon_t$ which has $E[y_t] = ct, \text{var}(y_t) = \sigma_e^2$

Stochastic trends: One that changes each run due to the random component of the process, such as $y_t = c + y_{t-1} + \varepsilon_t$ which has $\text{var}(y_t) = t\sigma_e^2$

AR models for non-stationary time series

$\{\Delta Y_t\}$ is AR(p) iff $\{Y_t\}$ is AR$(\text{p} + 1)$ with $\sum_{i=1}^{p} \beta_i = 1$

Hence if ΔY_t follows an AR(p) then $\{Y\}$ is not stationary. What is the implication of this?

- Previously we have been differencing $\{Y\}$ and fitting a (potentially stationary) AR(p) model to $\{\Delta Y_t\}$, but we don't have to do that!
- We can just fit an AR(p+1) model to $\{Y_t\}$.

Note that the claim is **not** being made here that if we have a non-stationary AR(p) model of $\{Y_t\}$ then if we difference and fit an AR(p-1) model to $\{\Delta Y_t\}$ then it will be stationary. It is not necessarily the case that $\{Y_t\}$ is $I(1)$ and hence we might have to difference multiple times to get to stationarity. The claim is merely that this relationship between differencing and AR(p) models holds, which might prove useful so that ***when*** we do have non-stationary ***and*** $I(1)$ series like $\{Y_t\}$, rather than differencing we might just be able to fit an AR(p-1).

Unit Roots or Difference

So which should we do? Use the Unit Roots or Difference?

- We know that $\{\Delta Y_t\}$ is AR(p-1) iff $\{Y_t\}$ is AR(p) with $\sum_{i=1}^{p} \beta_i = 1$ so if we have a unit root we can just difference it and get a not unit root
- But the parameters of both are consistently estimated by OLS.
- Problem with unit roots is that CLTs do not apply, hence we get non-standard inference as some regression estimates are not asymptotically normal. -Unit root is also more likely to be biased.
- Hence if we know we have a unit root it is preferable to difference it.
- But don't over-difference because that is also bad.

AR(1) Unit Roots

AR Representation

$$Y_t = \beta_0 + Y_{t-1} + u_t$$

MA Representation

$$Y_t = t\beta_0 + Y_0 + \sum_{i=1}^{t} u_j$$

Proof: Recursive Substitution

$$\begin{aligned}
&Y_t = \beta_0 + Y_{t-1} + u_t \\
&Y_t = \beta_0 + (\beta_0 + y_{t-2} + u_{t-1}) + u_t \\
&Y_t = 2\beta_0 + (\beta_0 + Y_{t-3} + u_{t-2}) + u_{t-1} + u_t \\
&... \\
&Y_t = h\beta_0 + Y_{t-h} + u_{t-h+1} + u_{t-h+t} \\
&Y_t = h\beta_0 + Y_{t-h} + \sum_{j=0}^{h-1} u_{t-j} \\
&\text{Then set } h = t \text{ and so} \\
&Y_t = t\beta_0 + Y_0 + \sum_{j=0}^{t-1} u_{t-j} = t\beta_0 + Y_0 + \sum_{j=1}^{t} u_j
\end{aligned}$$

Properties

(1) $t\beta_0$ is a deterministic function of time (process either grows or decays).

(2) $U_t = \sum_{j=1}^{t} u_j$ wanders randomly,

- $U_{t+1} = \sum_{j=1}^{t+1} u_j = \sum_{j=1}^{t} u_j + u_{t+1} = U_t + u_t + 1$ - highly persistent, new value is equal to previous plus unforecastable innovation.
- $Var(U_t) = t\sigma_u^2$ - clearly nonstationary.

AR(p) Unit Roots

AR Representation

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-1} + u_1 \ , \ \sum_{i=1}^{p} \beta_i = 1 \ , \ E\left[u_t \mid y_{t-1}\right] = 0$$

MA Representation

$$Y_t = t\beta_0 + Y_0 + V_t$$

- Where $V_t = \sum_{j=1}^{t} v_j$ and $\{v_t\}$ is stationary: a stochastic trend
  - If not then $\{v_t\}$ is serially correlated, but we still get random wandering behaviour.
  - If $\{v_t\}$ is serially uncorrelated then we have a random walk
    - Random walk is a special case of stationarity

Unit root (AR) process:

- $\{Y_t\}$ follows an AR(p) model with $\sum_{i=1}^{p} \beta_i = 1$
  - special case: $\beta_1 = 1$ in the AR(1) model
- Decomposes into sum of a deterministic trend (if $\beta_0 \neq 0$), a stochastic trend, and an initial value

$$Y_t = \beta_0 t + \sum_{s=1}^{t} v_t + Y_0$$

Testing for Unit Roots

Hypotheses under AR(1)

Model (constant only)

$$Y_t = \beta_0 + Y_{t-1} + u_t \ , \ E\left[u_t \mid y_{t-1}\right]$$

Unit root implies $\beta_1 = 1$, stationarity implies $\beta_1 < 1$, hence

$$H_0 : \beta_1 = 1 \quad H_1 : \beta_1 < 1$$

Equivalent form,

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t, E\left[u_t \mid y_{t-1}\right]$$

$$\begin{aligned} \Delta Y_t &= \beta_0 + \beta_1 Y_{t-1} - Y_{t-1} + u_t \\ &= \beta_0 + \left(\beta_1 - 1\right) Y_{t-1} + u_t \\ &= \beta_0 + \delta Y_{t-1} + u_t \end{aligned}$$

$$\delta = \beta_1 - 1 \ , \ \hat{\delta} = \hat{\beta}_1 - 1$$

$$\begin{aligned} &H_0 : \delta = 0 \\ &H_1 : \delta < 0 \end{aligned}$$

So a unit root corresponds to $\delta = 0$ while stationarity corresponds to $\delta < 0$.

Hypotheses under AR(p)

Model (constant only)

$$\Delta Y_t = \beta_0 + \delta Y_{t-1} + \sum_{i=1}^{p} \gamma_i \Delta Y_{t-i} + u_t$$

Model (constant & trend)

$$\Delta Y_t = \beta_0 + \alpha t + \delta Y_{t-1} + \sum_{i=1}^{p} \gamma_i \Delta Y_{t-i} + u_t$$

AR(2) (constant only) case:

$$\Delta Y_t = \beta_0 + \delta Y_{t-1} + \gamma_1 \Delta Y_{t-1} + u_t$$

$$\begin{aligned} Y_t - Y_{t-1} &= \beta_0 + \delta Y_{t-1} + \gamma_1 \left(Y_{t-1} - Y_{t-2}\right) + u_t \\ Y_t &= \beta_0 + \left(\delta + \gamma_1 + 1\right) Y_{t-1} - \gamma_1 Y_{t-2} + u_t \\ &= \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + u_t \end{aligned}$$

$$\beta_1 + \beta_2 = 1 \Rightarrow \left(\delta + \gamma_1 + 1\right) - \gamma_1 = \delta + 1 = 1 \Rightarrow \delta = 0$$

Hence AR(p) is same as AR(1) case. That is a unit root corresponds to $\delta = 0$ while stationarity corresponds to $\delta < 0$

ADF Test: Constant Only

$$\Delta Y_t = \beta_0 + \delta Y_{t-1} + \sum_{i=1}^{p} \gamma_i \Delta Y_{t-i} + u_i$$

(1) Hypotheses: $H_0 : \delta = 0 \quad H_1 : \delta < 0$

- Null implies $\{\Delta Y_t\}$ is AR(p), hence $\{Y_t\}$ unit root
  - NOTE: this doesn't imply that ΔY_t is stationary, we would need to test this separately, accepting the null only implies that $\{Y_t\}$ has a unit root.
- Alternative implies $\{Y_t\}$ is stationary AR(p+1).

(2) Test-statistic: $t = \frac{\hat{\delta}}{\text{s.e. }(\hat{\delta})} \xrightarrow{D} DF_{cn}$

(3) Decision Rule: reject if $t < c_\alpha$

(4) Critical Values: from DF CONSTANT distribution

Left-tail critical value	10%	5%	1%
$N[0,1]$	-1.28	-1.64	-2.33
DF_{cn} : constant only	-2.57	-2.86	-3.43

ADF Test: Constant and Trend

$$\Delta Y_t = \beta_0 + \alpha t + \delta Y_{t-1} + \sum_{i=1}^{p} \gamma_i \Delta Y_{t-i} + u_t$$

(1) Hypotheses: $H_0 : \delta = 0 \quad H_1 : \delta < 0$ -Null implies $\{\Delta Y_t\}$ is AR(p), hence $\{Y_t\}$ unit root - NOTE: this doesn't imply that ΔY_t is stationary, we would need to test this separately, accepting the null only implies that $\{Y_t\}$ has a unit root.

- Alternative implies $\{Y_t\}$ is trend stationary AR(p+1).
  - It is stationary about the trend

(2) Test-statistic: $t = \frac{\hat{\delta}}{\text{s.e. }(\hat{\delta})} \xrightarrow{D} DF_{tr}$

(3) Decision Rule: reject if $t < c_\alpha$

(4) Critical Values: from DF TREND distribution

Left-tail critical value	10%	5%	1%
$N[0,1]$	-1.28	-1.64	-2.33
DF_{cn} : constant only	-2.57	-2.86	-3.43
DF_{tr} : constant and trend	-3.21	-3.41	-3.96

Orders of Integration

If we conclude that $\{Y_t\}$ has a unit root (accept $H_0 : \delta = 0$) it **does not necessarily follow** that $\{\Delta Y_t\}$ **is stationary**.

$$\Delta Y_t = \beta_0 + \delta Y_{t-1} + \sum_{i=1}^{p} \gamma_i \Delta Y_{t-i} + u_t$$

Even if $\delta = 0$, it is still possible that $\sum_{i=1}^{p} \gamma_i = 1$

Hence $\{\Delta Y_t\}$ may itself have a unit root.

The **Order of integration of** $\{Y_t\}$ is the smallest $d = \{0, 1, 2, \ldots\}$ such that $\left\{\Delta^d Y_t\right\}$ is stationary, denoted $Y_t \sim I(d)$

Estimation

We estimate $I(d)$ by sequential ADF tests:

(1) Test for Unit Root in $\{Y_t\}$: If reject, then $d = 0$, else...

(2) Test for Unit Root in $\{\Delta Y_t\}$: If reject, then d = 1, else...

(3) Test for Unit Root in $\left\{\Delta^2 Y_t\right\} \ldots$

   - Where $\Delta^2 Y_t = \Delta\left(\Delta Y_t\right) = \Delta Y_t - \Delta Y_{t-1}$

I(0) : stationary processes

I(1): have stochastic trends, but differences are stationary

I(2) : randomly wandering, even more persistent (smoother) than I(1), but difference twice for stationarity.

We can reasonably approximate:

- I(0) processes by stationary AR models.
- I(1) processes by unit root AR models.

Spurious Regression vs Co-integration

Spurious Regression

Are $\{X_t\}$ and $\{Y_t\}$ ‘related’ in any causal or predicative sense?

- If $\{X_t\}$ and $\{Y_t\}$ are stationary and independent, then OLS consistently estimates a coefficient of zero on the X’s (hence they are not ‘related’)

When stochastic trends are present, this goes awry however…

- Suppose $\{X_t\}$ and $\{Y_t\}$ are independent random walks

$$X_t = \sum_{s=1}^{t} \varepsilon_{x,s}$$
$$Y_t = \sum_{s=1}^{t} \varepsilon_{y,s}$$

- With mean zero iid innovations (the epsilons are independent and identically distributed), and suppose that both are I(1) process with no deterministic drift.
- Despite this *OLS regression has a systematic tendency to find a statistically significant* relationship between $\{X_t\}$ and $\{Y_t\}$
- Given that $\{X_t\}$ and $\{Y_t\}$ are independent this significant is **spurious**.

Spurious regression: the systemic tendency to find statistically significant regression relationships between unrelated I(1) series.

Explanation

Can be due to having the same deterministic drift, but even if they don’t spurious regression still occurs, usually due to stochastic trends - ‘Random wandering with no tendency to revert to a fixed mean’.

Because stochastic trends exhibit long swings of increase and decline even unrelated I(1) series will tend to have periods in which they move in the same direction purely by chance.

Finding Spurious Regression

(1) Check that $\{X_t\}$ and $\{Y_t\}$ are I(1)

(2) Analyse the residuals from $\hat{u}_t = Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_t$

If $\{X_t\}$ and $\{Y_t\}$ are I(1) and unrelated then $\hat{u}_t$ will inherit stochastic trend.

Then $\hat{u}_t$ will be highly serially correlated and will look like an I(1) process.

Cointegration

We can, however, regress I(1) series on one another if they share common stochastic (and deterministic) trends.

$\{X_t\}$ and $\{Y_t\}$ **are cointegrated** if $\{X_t\}$ and $\{Y_t\}$ **are I(1)**, and **there exists a** θ (cointegration coefficient) such that $Y_t - \theta X_t \sim I(0)$

This arises because $\{X_t\}$ and $\{Y_t\}$ share a common stochastic (and deterministic) trend!

Mathematical illustration

$$Y_t = \sum_{s=1}^{t} (\mu + v_s) + w_{y,t} \ , \ X_t = \frac{1}{\theta} \sum_{s=1}^{t} (\mu + v_s) + w_{x,t}$$

Where $Y_t, X_t \sim I(1)$: both have stochastic trends, and $\Delta Y_t = \mu + v_t + w_{y,t} - w_{y,t-1} \sim I(0) =$ (their differences are stationary!)

$$\begin{aligned} Y_t - \theta X_t &= \left[\sum_{s=1}^{t} (\mu + v_s) + w_{y,t} \right] - \theta \left[\frac{1}{\theta} \sum_{s=1}^{1} (\mu + v_s) + w_{x,t} \right] \\ &= w_{y,t} - w_{x,t} \sim I(0) \end{aligned}$$

Implications for OLS

OLS makes sense for co-integration!

$$min \sum_{t=1}^{T} (Y_t - a - cX_t)^2$$

$\hat{\theta}$ minimises SSR since it removes the trend (approximately).

$$\hat{\theta} \xrightarrow{P} \theta$$

$\hat{\xi}_t = Y_t - \widehat{\theta} X_t \simeq Y_t - \theta X_t \sim I(0)$

Testing for Cointegration:

(1) Perform ADF test to verify that $Y_t, X_t \sim I(1)$

(2) If we know the cointegrating coefficient θ

- (i) Perform ADF test on $\xi_t = Y_t - \theta X_t$
- (ii) Null of unit root implies $\xi_t \sim I(1)$ hence no co-integration
- (iii) Reject null then conclude that $\xi_t \sim I(0)$ hence co-integrated

(3) If we don't know the cointegrating coefficient θ

- (i) Estimate θ by OLS, hence compute $\widehat{\xi_t} = Y_t - \widehat{\theta} X_t$
- (ii) Perform ADF test on $\widehat{\xi}_,$, USE ENGLE-GRANGER CRITICAL VALUES
- (iii) If reject null that $\xi_t \sim I(1)$ then conclude co-integrated

Dynamic Causal Effects

Distributed Lag Regression

Will a regression of Y_t on X_t consistently estimate a causal effect?

- We need Y_t on X_t to be stationary - though perhaps after transformations.
- If they are $I(1)$ and cointegrated then look at cointegration section.
- We have to ask the usual question: Does OR hold?
  - Is $\operatorname{cov}(X_t, u_t) = 0$?
    - What other determinants of Y might be correlated with X?
    - Is there any simultaneity/reverse causality?
- We MUST include other lags of X itself, since X is correlated with its lags (serially correlated) hence failing to do so would be an obvious cause of OVB (Omitted Variable Bias).

$$Y_t = \beta_0 + \gamma_0 X_t + \gamma_1 X_{t-1} + \gamma_2 X_{t-2} + \ldots + \gamma_r X_{t-r} + u_t$$

  - Notice objective here is to estimate causal effect, not to forecast.

Causal interpretation of parameters as dynamic multipliers:

$$\frac{\partial Y_{t+h}}{\partial X_t} = \begin{cases} \gamma_h \text{ if } h \in \{0, 1, \ldots, r\} \\ 0 \text{ otherwise} \end{cases}$$

γ_0 : impact effect

- The effect X_t has on Y_t immediately

γ_h : dynamic multiplier

- The effect X_t has on Y_t in h periods time
- Or the effect X_t h periods ago has had on Y_t

Cumulative dynamic multipliers:

$$\delta_h = \gamma_0 + \gamma_1 + \ldots + \gamma_h = \sum_{i=0}^{h} \gamma_i$$

The cumulative effect of X, summing over today and next h periods.

Equivalent to effect of a ‘permanent’ change in X, lasting $h + 1$ periods.